



Proceedings of the 2017 conference on Big Data from Space (BiDS'17)

28th - 30th November 2017
Toulouse (France)

Edited by P. Soille and P.G. Marchetti



The European Commission's
science and knowledge service
Joint Research Centre



This publication is a Conference report published by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication.

Contact information

Name: Pierre Soille
Address: European Commission, Joint Research Centre
Via Enrico Fermi 2749, TP 267, I-21027 Ispra (VA), Italy
Email: Pierre.Soille@ec.europa.eu
Tel.: +39 0332 78 9111

JRC Science Hub

<https://ec.europa.eu/jrc>

JRC108361

EUR 28783 EN

PDF ISBN 978-92-79-73527-1 ISSN 1831-9424 doi:10.2760/383579

Luxembourg: Publications Office of the European Union, 2017

© European Union, 2017

Reuse is authorised provided the source is acknowledged. The reuse policy of European Commission documents is regulated by Decision 2011/833/EU (OJ L 330, 14.12.2011, p. 39).

For any use or reproduction of photos or other material that is not under the EU copyright, permission must be sought directly from the copyright holders.

How to cite these proceedings: P. Soille and P.G. Marchetti (Eds.), *Proceedings of the 2017 conference on Big Data from Space. BIDS' 2017*, EUR 28783 EN, Publications Office of the European Union, Luxembourg, 2017, ISBN 978-92-79-73527-1, doi:10.2760/383579, JRC108361

Preface

Big Data from Space refers to massive spatio-temporal Earth and Space observation data collected by space-borne and ground-based sensors as well as the synergetic use of data coming from other sources and communities. Whether for Earth or Space observation, they qualify being called 'big data' given the sheer volume of sensed data (archived data reaching the exabyte scale), their high velocity (new data are acquired almost on a continuous basis and with an increasing rate), their variety (data are delivered by sensors acting over various frequencies of the electromagnetic spectrum in passive and active modes), as well as their veracity (sensed data are associated with uncertainty and accuracy measurements). Last but not least, the value of Big Data from Space depends on our capacity to extract information and meaning from them.

Big data from Space is undergoing sharp development with numerous new initiatives and breakthroughs from intelligent sensors to data science. These developments are empowering new approaches and applications in various and diverse domains. Information based developments such as spatio-temporal and long time-series data analytics, data cubes, smart data management, information extraction technologies, computational intelligence, platforms and services are undergoing a fast evolution. In addition, the recent multiplication of initiatives offering open access to Big Data from Space is giving momentum to the field by widening substantially the spectrum of users as well as awareness among the public while offering new opportunities for scientists, value-adding companies, and institutions. Benefits are expected at all levels from individual citizens to the whole society.

The overall objective of the BiDS conference cycle is to stimulate interactions and bring together researchers, engineers, users, infrastructure and service providers, interested in exploiting Big Data from Space.

Following the success of the 2014 and 2016 conferences on Big Data from Space held at the ESA in Frascati (Italy) and the Auditorio de Tenerife (Spain) respectively, the 2017 edition (BiDS'17) is again co-organised by the European Space Agency (ESA), the Joint Research Centre (JRC) of the European Commission, and the European Union Satellite Centre (SatCen). This edition is hosted by the French Centre National d'Etudes Spatiales (CNES) from 28 to 30 November 2017 in Toulouse, one of the key European cities with activities focused on space and aerospace developments and applications.

Since the last edition of the BiDS conference in March 2016, the Sentinel-1A and Sentinel-2A satellites have been complemented by three additional Copernicus Sentinel satellites: Sentinel-1B, Sentinel-2B, and Sentinel-5P launched on 22 April 2016, 7 March 2017, and 13 October 2017 respectively. It follows that the two first Sentinel missions are now almost in full operational mode with orbit repeat time of 6 and 5 days respectively. Sentinel-3A launched on 16 February 2016 has been handed for operations in July 2016 and the launch of Sentinel-3B is scheduled for 2018. Copernicus with its fleet of Sentinel satellites is the world's largest single Earth observation programme. Copernicus is directed by the European Commission in partnership with ESA that is responsible for the space component of the programme. Together with numerous other programmes and initiatives in Europe and beyond, free and open data are contributing to answering major societal questions such as those related to the environment, climate change, crisis management, and sustainable development goals. Regarding big data from space observations, the first catalogue of more than a billion stars from ESA's Gaia satellite was published in September 2016. This catalogue represents the largest all-sky survey of celestial objects to

date.

The focus of BiDS'17 is on the whole data life cycle, ranging from data acquisition by space borne and ground-based sensors to data management, analysis and exploitation in the domains of Earth Observation, Space Science, Solar System Objects, Space Situational Awareness, Secure Societies, etc. Special emphasis is put on highlighting synergies and cross-fertilisation opportunities. The main objectives of BiDS'17 include:

- Present and discuss results, progresses and emerging challenges, including open data access and open science;
- Involve users and service providers in the definition of a roadmap for future developments;
- Identify priorities for research, technology development and innovation;
- Widen competences and expertise of universities, research institutes, labs, SMEs and other industrial actors;
- Foster networking of experts and users towards better access and sharing of data, tools, infrastructures, resources and incubation of services;
- Leverage innovation, spin-in and spin-off of technologies, and business development arising from research and industry progress;
- Increase and promote the value stemming from the huge quantity of data made available at an ever increasing rate.

The presentations, discussions, and contacts established during the conference as well as the materials presented in these proceedings are contributing to these goals. A total of 183 papers were submitted for presentation at the conference. Following the peer-review process by members of the conference programme committee, 80 papers were selected for oral presentation. They are complemented by 46 poster presentations (for a total of 613 distinct co-authors with affiliations distributed over 23 different countries from all continents). Given the large number of presentations, the 3 day conference programme had to be structured around 9 pairs of parallel sessions complemented by the poster session. An additional demonstration/industry session (with 15 demos) has been organised alongside the poster session, to give actors the possibility to present live demos on big data from space applications.

The conference opening session was devoted to a series of enlightening institutional talks by CNES, ESA, SatCen, and the European Commission with the following speakers: Geneviève Campan (Centre National d'Études Spatiales), Andreas Veispak (European Commission, DG GROW), Giuseppe D'Amico (European Union Satellite Centre), Nicolaus Hanowski (European Space Agency), and Charles Macmillan (European Commission, DG JRC).

These proceedings consist of a collection of 126 short papers corresponding to the oral and poster presentations delivered at the conference. They are organised in sections matching the order of the conference sessions followed by the contributions that were presented during the poster session, also organised by topics. They provide a snapshot of the current research activities, developments, and initiatives in Big Data from Space. Further to the oral and poster contributions, the conference has been enriched by 5 enlightening invited keynote lectures addressing various big data topics of interest to Big Data from Space:

1. *Big Data management and Big Data mining as a service at CERN*
by Massimo Lamanna (European Organization for Nuclear Research, Geneva, Switzerland)
2. *Big Data? Small Data? Open Data!*
by Christoph Bruch (Helmholtz Open Science Coordination Office, Berlin, Germany)
3. *EO Ground Segment in China: an overview*
by Xiao Chen (Beijing Normal University, Beijing, China)
4. *Big Data at NASA*
by Lewis John McGibbney (National Aeronautics and Space Administration, USA)
5. *The stakes and prospects of Data Driven Modeling at CERFACS*
by Olivier Thual (Institut National Polytechnique de Toulouse, France)

As a tradition, the BiDS conference hosts contributions addressing the whole life-cycle of Big Data from Space, from data valorisation and preservation to on board processing, down to image and data analysis, including downstream exploitation. While a continued number of contributions are devoted to infrastructures and platforms enabling to exploit the value behind the volume, velocity, and variety of Big Data from Space, this third edition of the Big Data from Space conference shows a sharp increase of applications particularly related to large scale analysis including the temporal dimension in view of better understanding the dynamics of the processes that are shaping our planet and our universe. Other new trends regard information extraction using advanced machine learning techniques such as those based on deep learning and convolutional neural networks. The development of new standards to ensure the interoperability of Big Data from Space is also gaining attention similarly to data cubes and multidimensional array representations. All these topics as well as other generic key aspects of big data are mirrored onto dedicated sections in these proceedings. It is worth mentioning that few research initiatives presented in these proceedings are devoted to the collection and exploitation of in situ and crowd-sourced data as well as to the definition and set-up of large reference annotated datasets. It is expected that these topics will undergo increased attention in the near future. They would also definitely benefit from the support of dedicated research and institutional programmes.

Additional conference materials such as electronic version of the slides presented at the conference, including those regarding the opening session talks and keynote lectures, are available on the conference website: www.bigdatafromspace2017.org.

A great thanks goes to all authors and presenters of BiDS' 17 as well as the numerous participants (over 600 registrations spanning 48 countries over all continents). Together, they have ensured the success of the 2017 conference on Big Data from Space. A special thank goes to the Programme Committee members and the additional reviewers for their thorough reviews and detailed comments that were taken into account by the authors when preparing the final version of their paper included in these proceedings.

This edition of the BiDS conference is deeply grateful to CNES for its great support in having BiDS' 17 hosted in Toulouse.

Pierre Soille and Pier Giorgio Marchetti

Conference Chairs

General Chair

Pierre Soille European Commission, Joint Research Centre (JRC)

Co-Chairs

Sergio Albani European Union Satellite Centre (SatCen)
Jean-Pierre Gleyzes Centre National d'Études Spatiales (CNES), France
Pier Giorgio Marchetti European Space Agency (ESA), ESIN, Italy

Organising Committee

Sergio Albani European Union Satellite Centre (SatCen)
Simon Baillarin Centre National d'Études Spatiales (CNES), France
Peter Baumann Jacobs University Bremen, Germany
Lorenzo Bruzzone University of Trento, Italy
Mihai Datcu Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Marco Freire European Space Agency (ESA), ESTEC, The Netherlands
Jutta Graf Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Michele Iapaolo European Space Agency (ESA), ESRIN, Italy
Pier Giorgio Marchetti European Space Agency (ESA), ESRIN, Italy
Vicente Navarro European Space Agency (ESA), ESAC, Spain
Pierre Soille European Commission, Joint Research Centre (JRC)
Juan Luis Valero European Union Satellite Centre (SatCen)

Local Organising Committee

Simon Baillarin Centre National d'Études Spatiales (CNES), France
Laurence Amen Centre National d'Études Spatiales (CNES), France
Masse Antoine Centre National d'Études Spatiales (CNES), France

Programme Committee

Selim Aksoy Bilkent University, Turkey
Sergio Albani European Union Satellite Centre (SatCen)
Philippe Armbruster European Space Agency (ESA), ESTEC, The Netherlands
Christophe Arviset European Space Agency (ESA), ESAC, Spain
Simon Baillarin Centre National d'Études Spatiales (CNES), France
Peter Baumann Jacobs University Bremen, Germany
Francesca Bovolo Fondazione Bruno Kessler, Italy

Lorenzo Bruzzone	University of Trento, Italy
Francesco Casu	IREA, National Research Council (CNR), Italy
Esther Conway	Science and Technology Facilities Council, UK
Christina Corbane	European Commission, Joint Research Centre (JRC)
Mihai Datcu	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Yves-Louis Desnos	European Space Agency (ESA), ESRIN, Italy
Liang Feng	The University of Edinburgh, UK
Marco Freire	European Space Agency (ESA), ESTEC, The Netherlands
Steffen Fritz	International Institute for Applied Systems Analysis (IIASA), Austria
Paolo Gamba	University of Pavia, Italy
Hinnerk Gildhoff	SAP SE, Germany
Jean-Pierre Gleyzes	Centre National d'Études Spatiales (CNES), France
Jose Gomez-Dans	NCEO/UCL, UK
Jutta Graf	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Jacopo Grazzini	European Commission, Eurostat
Steve Groom	IPAC/Caltech, USA
Michele Iapaolo	European Space Agency (ESA), ESRIN, Italy
Jordi Inglada	Centre National d'Études Spatiales (CNES)–CESBIO, France
Francois Jocteur-Monrozier	Centre National d'Études Spatiales (CNES), France
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Doris Klein	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Riccardo Lanari	IREA, National Research Council (CNR), Italy
Henri Laur	European Space Agency (ESA), ESRIN, Italy
Samantha Lavender	Pixalytics Ltd, UK
Michele Lazzarini	European Union Satellite Centre (SatCen)
Jacqueline Le Moigne	National Aeronautics and Space Administration (NASA), USA
Sébastien Lefèvre	Université de Bretagne Sud, France
Gianluca Luraschi	European Maritime Safety Agency (EMSA)
Michele Manunta	IREA, National Research Council (CNR), Italy
Pier Giorgio Marchetti	European Space Agency (ESA), ESRIN, Italy
Pierre-Philippe Mathieu	European Space Agency (ESA), ESRIN, Italy
Katrin Molch	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Vicente Navarro	European Space Agency (ESA), ESAC, Spain
Osamu Ochiai	Japan Aerospace Exploration Agency (JAXA), Japan
Simon Oliver	Geoscience Australia
Andrea Patrono	European Union Satellite Centre (SatCen)
Peter Reinartz	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Sven Schade	European Commission, Joint Research Centre (JRC)
Michael Schick	EUMETSAT
John Schnase	National Aeronautics and Space Administration (NASA), USA
Jose Sobrino	Universidad de Valencia, Spain
Pierre Soille	European Commission, Joint Research Centre (JRC)
Peter Strobl	European Commission, Joint Research Centre (JRC)
Vasileios Syrris	European Commission, Joint Research Centre (JRC)
Corina Vaduva	University Politehnica of Bucharest, Romania
Raffaele Vitulli	European Space Agency (ESA), ESTEC, The Netherlands
Wolfgang Wagner	Vienna University of Technology, Austria
Gui-Song Xia	Wuhan University, China

Additional Reviewers

d'Angelo, Pablo	Deutsches Zentrum für Luft- und Raumfahrt (DLR), Germany
Geller, Gary	GEO Secretariat, Switzerland
Greidanus, Harm	European Commission, Joint Research Centre (JRC)
Hasenohr, Paul	European Commission, Joint Research Centre (JRC)
Lazzarini, Michele	European Union Satellite Centre (SatCen)
Lemoine, Guido	European Commission, Joint Research Centre (JRC)
Popescu, Anca	European Union Satellite Centre (SatCen)
Sabatino, Giovanni	European Space Agency (ESA)
Sixsmith, Joshua	Geoscience Australia

Table of Contents

<hr/>	
Big Data Processing I	
<hr/>	
INTERNET MAJOR ACTORS TECHNOLOGIES APPLIED TO METEOROLOGICAL SATELLITE: STREAM DATA APPLIED TO DATA PROCESSING	1
<i>Alain Montmory and Laure Chaumat</i>	
FROM HADOOP MAP/REDUCE TO SPARK: GAIA DPCC PERFORMANCE USE CASE	5
<i>Guillaume Eynard-Bontemps, Olivier Melet, Hugo Palacin, Florent Ventimiglia and Louis Noval</i>	
BIG DATA PROCESSING USING THE EODC PLATFORM	9
<i>Stefano Elefante, Vahid Naemi, Senmao Cao, Iftikhar Ali, Tuan Sy Le, Wolfgang Wagner and Christian Briese</i>	
<hr/>	
Machine Learning and Semantic Querying	
<hr/>	
SEMANTIC-SENSITIVE HASHING FOR CONTENT-BASED RETRIEVAL IN REMOTE SENSING IMAGES	13
<i>Thomas Reato, Begüm Demir and Lorenzo Bruzzone</i>	
SYSTEMATIC ESA EO LEVEL 2 PRODUCT GENERATION AS PRE-CONDITION TO SEMANTIC CONTENT-BASED IMAGE RETRIEVAL AND INFORMATION/KNOWLEDGE DISCOVERY IN EO IMAGE DATABASES	17
<i>Andrea Baraldi, Dirk Tiede, Martin Sudmanns and Stefan Lang</i>	
DATA CORRELATION: FUSING LOGFILES WITH PERFORMANCE COUNTERS TO DIAGNOSE PERFORMANCE ISSUES IN GROUND SYSTEMS	21
<i>Paschalis Veskos, Stathis Koukouvinos and Fabien Castel</i>	
AUTOMATIC DEBRIS DETECTION FOR SPACE SITUATIONAL AWARENESS BASED ON GPU TECHNOLOGY	25
<i>Francesco Diprima, Fabio Santoni, Fabrizio Piergentili, Vito Fortunato, Cristoforo Abbattista and Leonardo Amoroso</i>	
YOUNG TREE IDENTIFICATION USING MACHINE LEARNING ON SENTINEL-1 DATA ...	29
<i>Sandrine Daniel, Julian Klein, Johannes Petrat, Young Lee and Lee Brown</i>	
<hr/>	
Data Cubes and Multidimensional Arrays	
<hr/>	
THE SIX FACES OF THE DATA CUBE	32
<i>Peter Strobl, Peter Baumann, Adam Lewis, Zoltan Szantoi, Brian Killough, Matthew Purss, Max Craglia, Stefano Nativi, Alex Held and Trevor Dhu</i>	
DIGITAL EARTH AUSTRALIA – UNLOCKING INNOVATION AND CAPABILITY	36
<i>Trevor Dhu, David Gavin, David Hudson, Trent Kershaw, Adam Lewis, Leo Lymburner, Norman Mueller, Simon Oliver, Jonathon Ross, Andreia De Avila Siqueira and Medhavy Thankkapan</i>	
SENTINEL-1 DATA CUBE EXPLOITATION: TOOLS, PRODUCTS, SERVICES AND QUALITY CONTROL	40
<i>Iftikhar Ali, Vahid Naeimi, Senmao Cao, Stefano Elefante, Le Tuan Sy, Bernhard Bauer-Marschallinger and Wolfgang Wagner</i>	
GEOSPATIAL WEB SERVICES FOR BIG CLIMATE DATA: ON-DEMAND ACCESS TO AND PROCESSING OF ECMWF'S REANALYSIS DATA	44
<i>Julia Wagemann, Stephan Siemen and Sylvie Lamy-Thepaut</i>	
THE E-SENSING ARCHITECTURE FOR BIG EARTH OBSERVATION DATA ANALYSIS	48
<i>Gilberto Câmara, Gilberto Queiroz, Lubia Vinhas, Karine Ferreira, Ricardo Cartaxo Modesto Souza, Rolf Simoes, Eduardo Llapa, Luiz Fernando Assis and Alber Sanchez</i>	

Information Generation at Scale

MASS PROCESSING OF SENTINEL-1 AND LANDSAT DATA FOR MAPPING HUMAN SETTLEMENTS AT GLOBAL LEVEL 52
Christina Corbane, Martino Pesaresi, Panagiotis Politis, Vasileios Syrris, Aneta Florczyk J., Pierre Soille, Luca Maffenini, Armin Burger, Veselin Vasilev, Dario Rodriguez Aseretto, Filip Sabo, Lewis Dijkstra and Thomas Kemper

ON THE CONTRIBUTION OF 20 YEARS OF ATSR DATA AND GEODESIC P-SPLINE EFFICIENT SPATIAL SMOOTHING METHOD TO ITCZ TREND ANALYSIS..... 56
Elisa Castelli, Massimo Ventrucci, Fedele Greco, Massimo Valeri, Bianca Maria Dinelli, Enzo Papandrea and Stefano Casadio

QA4ECV: 35 YEARS OF DAILY ALBEDO BASED ON AVHRR AND GEO 59
Said Kharbouche, Jan-Peter Muller, Olaf Danne and Nadine Gobron

EXPLORING VEGETATION PHENOLOGY AT CONTINENTAL SCALES: LINKING TEMPERATURE-BASED INDICES AND LAND SURFACE PHENOLOGICAL METRICS 63
Raul Zurita Milla, Romulo Goncalves, Emma Izquierdo Verdiguier and Frank Ostermann

LARGE SCALE FLOOD RECURRENCE MAP USING SAR DATA 67
Marco Chini, Ramona Pelich, Renaud Hostache, Patrick Matgen, José Manuel Delgado Blasco and Giovanni Sabatino

Visualization

INTERACTIVE VISUALISATION AND ANALYSIS OF GEOSPATIAL DATA WITH JUPYTER 71
Davide De Marchi, Armin Burger, Pieter Kempeneers and Pierre Soille

WEBASSEMBLY FOR EO DATA VALORIZATION, RUNNING (LEGACY) TIME-CONSUMING PROCESSINGS IN THE BROWSER..... 75
Nicolas Decoster, Julien Gaucher and Julien Nosavan

INTEGRATION OF WEB WORLD WIND AND SENTINEL HUB - A GLOBAL 4D BIG DATA EXPLORATION AND COLLABORATION PLATFORM 79
Grega Milcinski, Guenther Landgraf, Patrick Hogan and Paulo Sacramento

Interoperability, Standards, and Regulation

OGC BIG DATA WHITE PAPER- EXTENDED ABSTRACT 83
Marie-Francoise Voidrot and George Percivall

SPACE BIG DATA, SMALL EARTH LAWS: OVERCOMING THE REGULATORY BARRIERS TO THE USE OF SPACE BIG DATA APPLICATIONS 86
Dimitra Stefoudi

STARE - TOWARD UNPRECEDENTED GEO-DATA INTEROPERABILITY 90
Kwo-Sen Kuo and Michael L Rilee

CREATING VIRTUAL SEMANTIC GRAPHS ON TOP OF BIG DATA FROM SPACE 94
Konstantina Bereta and Manolis Koubarakis

BIG DATA CHALLENGES IN GEOSS 98
Stefano Nativi, Joost van Bemmelen, Mattia Santoro and Guido Colangeli

Big Data Processing II

COPERNICUS AND AIS DATA FUSION AND INFORMATION MANAGEMENT FOR MARITIME TASKS – PRELIMINARY RESULTS 102
José Manuel Delgado Blasco, Claudio Manganiello, Pier Giorgio Marchetti, Massinno Marrazzo and Mauro Arcorace

LESSONS LEARNED OVER THE PAST THREE YEARS, ON THE BIGDATA USAGE FOR PROCESSING GAIA DATA IN CNES	106
<i>Frederic Pailler, Laurence Chaoul, François Riclet and Chantal Panem</i>	

BEYOND SENTINEL-2 WITH URTHEDAILY CONSTELLATION	110
<i>Jose Julio Ramos</i>	

Time Series

A NEW PARADIGM FOR THE EXPLOITATION OF THE SEMANTIC CONTENT OF LARGE ARCHIVES OF SATELLITE REMOTE SENSING IMAGES	114
<i>Lorenzo Bruzzone, Manuel Bertoluzza and Francesca Bovolo</i>	

DETECTING ABNORMAL EVENTS IN MULTIVARIATE TELEMETRIES THANKS TO COVARIANCE ANALYSIS	118
<i>Clémentine Barreyre, Béatrice Laurent, Jean-Michel Loubes, Bertrand Cabon and Loïc Boussoif</i>	

CLOUD APPROACH TO AUTOMATED CROP CLASSIFICATION USING SENTINEL-1 IMAGERY	122
<i>Andrii Shelestov, Mykola Lavreniuk, Andrii Kolotii, Vladimir Vasiliev, Leonid Shumilo and Nataliia Kussul</i>	

SPATIO-TEMPORAL ANALYSIS OF CHANGE WITH SENTINEL IMAGERY ON THE GOOGLE EARTH ENGINE	126
<i>Morton J. Canty and Allan A. Nielsen</i>	

OPERATIONAL APPLICATION OF THE FULL LANDSAT TIMESERIES TO SERVICE INDUSTRY IN THE AUSTRALIAN RANGELANDS	130
<i>Peter Scarth</i>	

Public and Private Platforms

PEPS – THE FRENCH COPERNICUS COLLABORATIVE GROUND SEGMENT	134
<i>Stephane Duprat, Camille Louge, Marc Ferrer, Vincent Garcia, Driss El Maalem, Mireille Paulin, Erwann Poupart and Christophe Taillan</i>	

EUMETSAT, ECMWF & MERCATOR OCÉAN PARTNERS DIAS	138
<i>Michael Schick, Martin Dillmann, Lothar Wolf, Joana Miguens and Miruna Stoicescu</i>	

ASB – A PLATFORM AND APPLICATION AGNOSTIC SOLUTION FOR IMPLEMENTING COMPLEX PROCESSING CHAINS OVER GLOBALLY DISTRIBUTED PROCESSING AND DATA RESOURCES	142
<i>Bernard Valentin, Matthieu Melcot, Leslie Gale, Philippe Mougnaud and Michele Iapaolo</i>	

ONEATLAS, AIRBUS DEFENCE AND SPACE DIGITAL PLATFORM FOR IMAGERY	146
<i>Laurent Gabet, Philippe Nonin, Salvador Cavadini, Mathias Ortner and Matthieu Rouget</i>	

PROBA-V MISSION EXPLOITATION PLATFORM	150
<i>Erwin Goor, Jeroen Dries and Dirk Daems</i>	

Analysing the Temporal Dimension

PRESERVATION AND HARMONIZATION OF HISTORICAL AVHRR LAC DATA TO SERVE THE NEEDS OF USERS IN CLIMATE RESEARCH	154
<i>Stefan Wunderle, Christoph Neuhaus, Fabia Huesler, Andrew Brooks, Neil Lonie, Mirko Albani, Sergio Folco and Rosemarie Leone</i>	

EXPLOITATION OF ENVISAT ASAR AND SENTINEL-1 SAR DATA IN SUPPORT OF CARBON AND WATER CYCLE STUDIES	157
<i>Maurizio Santoro, Oliver Cartus, Andreas Wiesmann, Urs Wegmüller, Josef Kellndorfer, Christiane Schmullius, Pierre Defourny, Olivier Arino, Marcus Engdahl and Frank Martin Seifert</i>	

SYSTEM FOR AUTOMATIZED SENTINEL-1 INTERFEROMETRIC MONITORING	161
<i>Milan Lazecky</i>	
THE VALUE OF SAR BIG DATA FOR GEOHAZARD APPLICATIONS: AUTOMATED GRID PROCESSING OF ERS-1/2 AND ENVISAT DATA IN ESA'S G-POD	165
<i>Francesca Cigna and Deodato Tapete</i>	
EXPLOITING OCEAN OBSERVATION AND SIMULATION BIG DATA TO IMPROVE SATELLITE-DERIVED GEOPHYSICAL PRODUCTS: ANALOG STRATEGIES	169
<i>Ronan Fablet, Phi Viet, Redouane Lguensat, Pierre-Henri Horrein and Bertrand Chapron</i>	
<hr/> Big Data Selection and Compositing <hr/>	
SENTINEL-2 DASHBOARD FOR SPATIO-TEMPORAL ANALYSIS OF GLOBAL SCENE COVERAGE	173
<i>Martin Sudmanns, Hannah Augustin, Anna-Maria Cavallaro, Dirk Tiede and Stefan Lang</i>	
OPTIMISING SENTINEL-2 IMAGE SELECTION IN A BIG DATA CONTEXT	177
<i>Pieter Kempeneers and Pierre Soille</i>	
SCALABLE CLOUD-BASED COMPUTATION OF CONSISTENT SURFACE REFLECTANCE MOSAICS AT 10M FROM SENTINEL-2 AND LANDSAT-8 MISSIONS	181
<i>Konstantinos Karantzalos and Athanasios Karmas</i>	
<hr/> Large Scale Data Management <hr/>	
A MODEL-DRIVEN BIG DATA ARCHITECTURE FOR PLANETARY DATA ARCHIVES AND RESEARCH	185
<i>Daniel Crichton, John S Hughes, Sean Hardman, Emily Law, Thomas Stein and Reta Beebe</i>	
LARGE SCALE DATA MANAGEMENT OF ASTRONOMICAL SURVEYS WITH ASTROSPARK	189
<i>Mariem Brahem, Karine Zeitouni and Laurent Yeh</i>	
CONSIDERING SCALE OUT ALTERNATIVES FOR BIG DATA VOLUME DATABASES WITH POSTGRESQL	193
<i>Pilar de Teodoro, Sara Nieto, Jesus Salgado and Christophe Arviset</i>	
ARCHIVE MANAGEMENT OF NASA EARTH OBSERVATION DATA TO SUPPORT CLOUD ANALYSIS	197
<i>Christopher Lynnes, Kathleen Baynes and Mark McInerney</i>	
ONLINE EARTH OBSERVATION DATA MANAGEMENT	201
<i>Nicolas Weiland, Stephan Kiemle, Markus Kunze and Torben Keßler</i>	
<hr/> Cloud Computing <hr/>	
ORGANIZING ACCESS TO COMPLEX MULTI-DIMENSIONAL DATA: AN EXAMPLE FROM THE ESA SEOM SINCOHMAP PROJECT	205
<i>Alexander Jacob, Fernando Vicente-Guijalba, Harald Kristen, Armin Costa, Bartolomeo Ventura, Roberto Monsorno and Claudia Notarnicola</i>	
LARGE SPATIAL SCALE GROUND DISPLACEMENT MAPPING THROUGH THE P-SBAS PROCESSING OF SENTINEL-1 DATA ON A CLOUD COMPUTING ENVIRONMENT	209
<i>Claudio De Luca, Manuela Bonano, Francesco Casu, Riccardo Lanari, Michele Manunta, Mariarosaria Manzo and Ivana Zinno</i>	
BIG DATA FROM ESA EARTHNET THIRD PARTY MISSION PROGRAMME: OPPORTUNITIES AND FUTURE EVOLUTION	211
<i>Giuseppe Ottavianelli, Mirko Albani, Roberto Biasutti, Bianca Hoersch, H�erv�e Jeanjean, Henri Laur and Bruno Schmitt</i>	

Data Preservation and Valorisation

ON THE SHOULDERS OF GIANTS: PROTOTYPING THE HERO VIRTUAL RESEARCH ENVIRONMENT FOR DATA VALORISATION OF HERITAGE MISSIONS	214
<i>Mirko Albani, Joost van Bemmelen and Giancarlo Rivolta</i>	
LONG-TERM DATA PRESERVATION DATA LIFECYCLE AND STANDARDISATION PROCESS	217
<i>Mirko Albani, Rosemarie Leone, Katrin Molch, Razvan Cosac and Iolanda Maggio</i>	
TOWARDS A PRESERVATION CONTENT STANDARD FOR EARTH OBSERVATION DATA .	221
<i>Hampapuram Ramapriyan, Dawn Lowe, Andrew Mitchell and Kevin Murphy</i>	
SPOT WORLD HERITAGE – SPOT 1-5 DATA CURATION AND VALORIZATION WITH NEW ENHANCED SWH PRODUCTS	225
<i>Julien Nosavan, Agathe Moreau, Antoine Masse, Benoît Chausserie-Laprée and Claire Caillet</i>	
20-YEARS OF ESA SPACE SCIENCE DATA ARCHIVES MANAGEMENT	229
<i>Christophe Arviset, Deborah Baines, Isa Barbarisi, Sebastien Besse, Guido De Marchi, Beatriz Martinez, Arnaud Masson, Bruno Merin, Jesus Salgado and Claire Vallat</i>	

Community platforms

CLOUD BASED EARTH OBSERVATION DATA EXPLOITATION PLATFORMS	233
<i>Antonio Romeo, Salvatore Pinto, Alessandro Marin and Sveinung Loekken</i>	
VIRTUAL EXPLOITATION ENVIRONMENT DEMONSTRATION FOR ATMOSPHERIC MISSIONS	236
<i>Stefano Natali, Simone Mantovani, Gerhard Triebnig, Daniel Santillan, Marcus Hirtl, Barbara Scherllin-Pirscher and Cristiano Lopes</i>	
FORESTRY TEP RESPONDS TO USER NEEDS FOR SENTINEL DATA VALUE ADDING IN CLOUD	239
<i>Tuomas Häme, Renne Tergujeff, Yrjö Rauste, Clive Farquhar, Peter van Zetten, Philip Kershaw, Arnaud De Groof, Jarno Hämäläinen, Joost van Bemmelen and Frank Martin Seifert</i>	
MONITORING URBANIZATION WITH BIG DATA FROM SPACE - THE URBAN THEMATIC EXPLOITATION PLATFORM	243
<i>Jakub Balhar, Thomas Esch, Hubert Asamer, Martin Boettcher, Enguerran Boissier, Andreas Hirner, Emmanuel Mathot, Mattia Marconcini, Annekatrin Metz, Hans Permana, Tomas Soukup, Soner Ureyen, Vaclav Svaton and Julian Zeidler</i>	
FAST MI-SAFE PLATFORM: FORESHORE ASSESSMENT USING SPACE TECHNOLOGY ..	247
<i>Joan Sala Calero, Gerrit Hendriksen, Jasper Dijkstra, Amrit Cado Van der Lelij, Mindert de Vries, Rudie Ekkelenkamp and Edward P. Morris</i>	

Object Detection, Hierarchical Image Segmentation, and Mosaicking

EFFICIENT AND LARGE-SCALE LAND COVER CLASSIFICATION USING MULTISCALE IMAGE ANALYSIS	251
<i>François Merciol, Balem Thibaud and Sébastien Lefèvre</i>	
UNSUPERVISED OBJECT DETECTION ON REMOTE SENSING IMAGERY USING HIERARCHICAL IMAGE REPRESENTATIONS AND DEEP LEARNING	255
<i>Kostas Stamatiou, Georgios Ouzounis and Nikki Aldeborgh</i>	
URBAN BASELINE CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORKS ON SENTINEL-2 IMAGES	259
<i>Maria Kesa, Eleni Kroupi, Victor Navarro, Camille Pelloquin, Bahaaeddin Alhaddad, Laura Moreno and Aureli Soria-Frisch</i>	
SPUSPO: SPATIALLY PARTITIONED UNSUPERVISED SEGMENTATION PARAMETER OPTIMIZATION FOR EFFICIENTLY SEGMENTING LARGE HETEROGENEOUS AREAS ..	263
<i>Stefanos Georganos, Tais Grippa, Moritz Lennert, Sabine Vanhuysse and Eléonore Wolff</i>	

A GLOBAL MOSAIC FROM COPERNICUS SENTINEL-1 DATA 267
Vasileios Syrris, Christina Corbane and Pierre Soille

Institutional Platforms

THE JRC EARTH OBSERVATION DATA AND PROCESSING PLATFORM 271
Pierre Soille, Armin Burger, Davide De Marchi, Paul Hasenohr, Pieter Kempeneers, Dario Rodriguez Aseretto, Vasileios Syrris and Veselin Vasilev

A PLATFORM FOR MANAGEMENT AND EXPLOITATION OF BIG GEOSPATIAL DATA
IN THE SPACE AND SECURITY DOMAIN 275
Sergio Albani, Michele Lazzarini, Paulo Nunes and Emanuele Angiuli

MUSCATE: A VERSATILE DATA AND SERVICES INFRASTRUCTURE COMPATIBLE
WITH PUBLIC CLOUD COMPUTING 279
Joelle Donadieu, Simon Baillarin, Marc Leroy, Robert Ngo, Julien Nosavan, Arnaud Selle, Céline L'Helguen, Roger Rutakaza Maneno, Thierry Segur, Bastien Julie and Laurent Favot

COMBINING SMALL HOUSEKEEPING DATA LAKES INTO A SHARED BIG DATA
INFRASTRUCTURE AT ESOC - ACHIEVEMENTS AND FUTURE EVOLUTION 283
Rui Santos, Gustavo Marques and James Eggleston

EVOLVING JASMIN: HIGH PERFORMANCE ANALYSIS AND THE DATA DELUGE 287
Neil Massey, Philip Kershaw, Matt Pritchard, Matt Pryor, Sam Pepler, Jonathan Churchill and Bryan Lawrence

Deep Learning and Neural Networks

SEMANTIC SEGMENTATION USING DEEP NEURAL NETWORKS FOR SAR AND
OPTICAL IMAGE PAIRS 289
Wei Yao, Dimitrios Marmanis and Mihai Datcu

ARTIFICIAL GENERATION OF BIG DATA FOR IMPROVING IMAGE CLASSIFICATION:
A GENERATIVE ADVERSARIAL NETWORK APPROACH ON SAR DATA 293
Dimitrios Marmanis, Wei Yao, Fathalrahman Adam, Mihai Datcu, Peter Reinartz, Konrad Schindler, Jan Dirk Wegner and Uwe Stilla

SEA LEVEL ANOMALY PREDICTION USING RECURRENT NEURAL NETWORKS 297
Anne Braakmann-Folgmann, Ribana Roscher, Susanne Wenzel, Bernd Uebbing and Jürgen Kusche

DEEP SELF-TAUGHT LEARNING FOR REMOTE SENSING IMAGE CLASSIFICATION 301
Anika Bettge, Ribana Roscher and Susanne Wenzel

FORECASTING IONOSPHERIC TOTAL ELECTRON CONTENT MAPS WITH DEEP
NEURAL NETWORKS 305
Noëlie Cherrier, Thibaut Castaings and Alexandre Boulch

Posters: Data Storage, Catalogues, and Data Management

CROSS-MATCH OF ASTROMETRIC CATALOGUES PERFORMED WITH DATABASE
SPATIAL TECHNOLOGIES 309
Angelo Fabio Mulone, Roberto Morbidelli, Rosario Messineo, Mario Lattanzi, Ruben De March and Alberto Vecchiato

SERVING CONTINUOUS AND GLOBAL HIGH RESOLUTION SATELLITE DATA - AN
EXAMPLE BASED ON SENTINEL-2 DATA 313
Rouven Volkmann, Christian Strobl, André Twele, Torsten Heinen and Christoph Reck

EFFICIENT PROTOCOLS TO STORE AND TRANSMIT FOR BIG DATA GENERATED BY
EARTH OBSERVATION SATELLITES 317
Yousuke Ikehata, Takahiro Minami, Yuji Shimomura, Hidekazu Mikai and Naoyuki Fujita

STORAGE OPTIMIZATION ACCORDING TO MISSION-ORIENTED CRITERIA	321
<i>Olivier Queyrut, Xavier Geoffret, Pierre-Marie Brunet, Patrick Ginet, Cyrille Parra and Denis Gutfreund</i>	
EUMETSAT SUBMISSION INFORMATION PACKAGE (SIP)	325
<i>David Berry and Michael Schick</i>	
<hr/> Posters: Computing environments <hr/>	
FROM HPC TO HYBRID CLOUD COMPUTING	328
<i>Sébastien Dorgan, Vincent Gaudissart and Stephan Aimé</i>	
DOCKER USED ON SPATIAL GROUND SEGMENT	332
<i>Christophe Baroux, Jean-Christophe Dislaire and Yanis Lisima</i>	
JUPYTEP IDE AS A CONCEPT OF INTEGRATED DEVELOPMENT ENVIRONMENT FOR EO DATA CLOUD-BASED PROCESSING SOLUTIONS	336
<i>Daniel Zinkiewicz, Jacek Rapiński and Michał Bednarczyk</i>	
EARTH OBSERVATION DATA EXPLOITATION IN THE ERA OF BIG DATA: ESA'S RESEARCH AND SERVICE SUPPORT ENVIRONMENT	340
<i>Roberto Cuccu, Giovanni Sabatino, José Manuel Delgado Blasco, Joost van Bemmelen and Giancarlo Rivolta</i>	
INTERCONNECTING PLATFORMS VIA WPS: EXPERIENCE FROM THE CTEP/PEPS CONNECTION	344
<i>Sebastien Clerc, Nicolas Gilles, Mireille Paulin, Giulio Ceriola, Emmanuel Poupart, Vincent Garcia, Christian Taillan, Michael Aspetsberger, Sylvie Barrau Huguet and Yoann Moreau</i>	
<hr/> Posters: Deep learning <hr/>	
RF SIGNAL CHARACTERIZATION USING DEEP LEARNING	348
<i>Ahmad Berjaoui and Adrien Elfassi</i>	
OFF THE SHELF DEEP LEARNING PIPELINE FOR REMOTE SENSING APPLICATIONS ...	352
<i>Rachit Tripathi, Adrien Chan-Hon-Tong and Alexandre Boulch</i>	
DEEP LEARNING FOR DENOISING OF SATELLITE IMAGES	356
<i>Pierre Blanc-Paques and Renaud Fraisse</i>	
<hr/> Posters: Data Integration and Fusion <hr/>	
DATA INTEGRATION OF REMOTE SENSING AND IN SITU DATA FROM SEVERAL SOLAR SPACE MISSIONS FOR SPACE WEATHER SERVICES	359
<i>Marta Casti, Silvano Fineschi, Rosario Messineo, Ester Antonucci, Angelo Fabio Mulone, Alessandro Bemporad, Andrea Fonti, Roberto Susino, Fabio Filippi, Daniele Telloni, Filomena Solitro, Gianalfredo Nicolini and Michele Martino</i>	
HOW TO APPROACH TO EARTHQUAKE PHYSICS STUDY BY AN INTEGRATED SATELLITE AND GROUND DATA ANALYSIS SYSTEM: THE SAFE ESA-FUNDED PROJECT	363
<i>Angelo De Santis, Cristoforo Abbattista, Lucilla Alfonsi, Leonardo Amoruso, Marianna Carbone, Claudio Cesaroni, Gianfranco Cianchini, Giorgiana De Franceschi, Anna De Santis, Rita Di Giovambattista, Daniela Drimaco, Alessandro Ippolito, Dedalo Marchetti, Francisco Jose Pavon Carrasco, Loredana Perrone, Alessandro Piscini, Luca Spogli and Francesca Santoro</i>	
HARNESSING BIG DATA FOR AGRICULTURE MONITORING: COMBINING REMOTE SENSING, OPEN ACCESS DATA AND CROWDSOURCING	367
<i>Raphaël d'Andrimont, Guido Lemoine, Marijn van der Velde and Christina Corbane</i>	

NEXT STEP FOR BIG DATA INFRASTRUCTURE AND ANALYTICS FOR THE SURVEILLANCE OF THE MARITIME TRAFFIC FROM AIS & SENTINEL SATELLITE DATA STREAMS	371
<i>Ronan Fablet, Nicolas Bellec, Laetitia Chapel, Chloé Friguet, René Garello, Pierre Gloaguen, Guillaume Hajduch, Sébastien Lefèvre, François Merciol, Pascal Morillon, Christine Morin, Mathieu Simonin, Romain Tavebard, Cédric Tedeschi and Rodolphe Vadaine</i>	
MERGING INSAR AND GNSS METEOROLOGY: HOW CAN WE MINE INSAR AND GNSS DATABASES TO EXTRACT AND VISUALIZE INFORMATION ON ATMOSPHERE PROCESSES?	375
<i>Giovanni Nico, Amaia Gil, Marco Quartulli, Pedro Mateus and Joao Catalao</i>	

Posters: Data Processing, Analytics, and Vizualisation

INNOVATIVE APPROACH FOR PMM DATA PROCESSING AND ANALYTICS	379
<i>Ruben De March, Maurizio Deffacis, Fabio Filippi, Andrea Fonti, Chiara Leuzzi, Marco Montironi, Angelo Fabio Mulone and Rosario Messineo</i>	
BENCHMARKING C++ IMAGE PROCESSING LIBRARIES FOR THE EUCLID SCIENCE GROUND SEGMENT	383
<i>Peter Kettig and Antoine Basset</i>	
A FRAMEWORK FOR OBJECT DETECTION IN SATELLITES IMAGES	387
<i>Mathias Ortner, Pierre Blanc-Paques, Ségolène Bourrienne, Laurent Gabet and Jean-François Faudi</i>	
SCOUTER: GEO-SOCIAL AND REAL-TIME ANOMALY CONTEXTUALIZATION	390
<i>Badre Belabess, Musab Bairat, Jérémy Lhez, Olivier Curé, Houda Khrouf and Gabriel Kepeklian</i>	
KNOWLEDGE RETRIEVAL STRATEGY FOR SATELLITES SYSTEM MONITORING BASED ON DATA ANALYTICS TECHNIQUES	394
<i>Carlo Ciancarelli, Arturo Intelisano and Silvio Giuseppe Neglia</i>	
ADAPTING EFFICIENTLY MID-LEVEL FEATURES TO HIGH RESOLUTION SATELLITE IMAGES INDEXING	398
<i>Assia Kourgli, Lynda Bouchemakh, Aichouche Belhadj-Aissa and Youcef Oukil</i>	
TOWARDS A MAP OF THE EUROPEAN TREE COVER BASED ON SENTINEL-2	402
<i>Thor-Bjørn Ottosen, Geoffrey Petch, Mary Hanson and Carsten Ambelas Skjøth</i>	
BIGEARTH-ACCURATE AND SCALABLE PROCESSING OF BIG DATA IN EARTH OBSERVATION	406
<i>Begüm Demir</i>	
NEW METHODOLOGIES TO ANALYZE BIG DATA FROM SPACE WITH A SPECIAL FOCUS ON EARTH OBSERVATION DATA	410
<i>László Bacsárdi, Gergely Bencsik and Zoltán Pödör</i>	
BIG DATA VISUALIZATION TOOLS IN EO MOBILE APPS	414
<i>Carla Orrù, Jakub Balhar, Adrian Stoica, Paulo Sacramento and Giancarlo Rivolta</i>	
BIG LUNAR DATA VISUALIZATION AND ANALYSIS	417
<i>Emily Law, George Chang, Richard Kim and Shan Malhotra</i>	

Posters: Time Series, Multitemporal Analysis, and Change Detection

THE REVISED TIME-FREQUENCY ANALYSIS (R-TFA) TOOL OF THE SWARM MISSION .	421
<i>Georgios Balasis, Constantinos Papadimitriou, Athanassios Daglis, Omiros Giannakis, Sigiava Giamini and Georgios Vasalos</i>	

BIG DATA ANALYTICS APPROACH FOR GEOSPATIAL INVESTIGATION OF URBAN GEOHAZARDS IN NAPLES, ITALY, WITH COSMO-SKYMED PERSISTENT SCATTERERS	425
<i>Deodato Tapete, Francesca Cigna, Pietro Milillo, Daniele Perissin, Carmine Serio and Giovanni Milillo</i>	
MULTITEMPORAL INTERFEROMETRY AND BIG DATA – CASE OF ALBANIA	429
<i>Neki Frasheri, Gudar Beqiraj and Salvatore Bushati</i>	
CLIMATE EXTREME EVENTS DETECTION BASED ON WEATHER FORECASTING VARIABLES COMBINATION	433
<i>Javier Lozano, Marco Quartulli and Igor G. Olaizola</i>	
MEDUSA: MULTITEMPORAL EARTH OBSERVATION DATAMASS FOR URBAN SPRAWL AFTERCARE	437
<i>Elise Koeniguer, Karine Adeline, Jérôme Besombes, Alexandre Boulch, Xavier Ceamenos, Adrien Chan-Hon-Tong, Guillaume Dufour, Fabrice Janez, Aurélie Michel, Aurélien Plyer, François Rogier and Pauline Trouvé-Peloux</i>	
BIG EARTH OBSERVATION DATA FOR FAST DETECTION OF DEFORESTATION USING ADAPTATIVE FILTERING	441
<i>Alber Sánchez and Gilberto Câmara</i>	
EO BIG DATA ANALYTICS FOR THE DISCOVERY OF NEW TRENDS OF MARINE SPECIES HABITATS IN A CHANGING GLOBAL CLIMATE.....	445
<i>Zoheir Sabeur, Gianluca Correndo, Galina Veres, Banafshe Arbab-Zavar, Geoffrey Neumann, Thomas Ivall, Fabien Castel, Jean-Michel Zigna and Jose Lorenzo</i>	
SOCIOECOLOGICAL CARBON PRODUCTION IN MANAGED AGRICULTURAL-FOREST LANDSCAPES	449
<i>Jiquan Chen, Kyla Dahlin, Ranjeet John, Gabriela Shirkey, Susie R. Wu, Phil Robertson, Steve Hamilton, Lauren Cooper, Dave Lusch, Arnon Karnieli, Raffaele Laforteza, Giovanni Sylos Labini and Angelo Amodio</i>	
<hr/> Posters: Application oriented platforms, services, and toolboxes <hr/>	
THE COASTAL WATERS RESEARCH SYNERGY FRAMEWORK, FOR UNLOCKING OUR POTENTIAL FOR COASTAL INNOVATION GROWTH.....	453
<i>Miguel Terra-Homem, Nuno Grosso, Nuno Catarino, Rory Scarrot, Eirini Politi and Abigail Cronin</i>	
OVERVIEW OF THE ESA ATMOSPHERIC DATA CENTER: EVDC	457
<i>Paolo Castracane, Angelika Dehn, Paul Kiernan, Shane Carty, Ann Mari Fjaeraa, Thomas Espe, Ian Boyd, Alastair McKinstry, Conor Delaney and Johannes Hansen</i>	
FOOD SECURITY – THEMATIC EXPLOITATION PLATFORM: BIG DATA FOR SUSTAINABLE FOOD PRODUCTION	461
<i>Heike Bach, Silke Migdall, Markus Muerth, Phillip Harwood, Andrea Colapicchioni, Antonio Cuomo, Sven Gilliams, Erwin Goor, Tom Van Roey, Andy Dean, Jason Suwala, Antonio Romeo, Esther Amler, Philippe Mougnaud and Espen Volden</i>	
EO4WILDLIFE: A CLOUD PLATFORM TO EXPLOIT SATELLITE DATA FOR ANIMAL PROTECTION	465
<i>Fabien Castel, Gianluca Correndo and Alan F. Rees</i>	
APPLICATION OF EARTH OBSERVATION TO A UGANDAN DROUGHT AND FLOOD MITIGATION SERVICE.....	469
<i>Samantha Lavender, Paul Healy, Ian Robinson, Regina Lally, Stephanie Ties, Darren Lumbroso, Elizabeth Valone, George Gibson, Lucrezia Tincani, Caroline Chambers, Chris Doyle, Andrew Lavender, Alexa Williams, John Auburn, Elma Jenkins, Arnaud Le Carvenec, Miguel Morgado, Simon Reid, Luca Innocente, Lisa Osborne, Heather Pitcher, Sebastian Clarke, Jamie Williams, Gina Tsarouchi, Jimmy Okori, Mark Harrison and Richard Jones</i>	

RHETICUS: A CLOUD-BASED GEO-INFORMATION SERVICE FOR GROUND INSTABILITIES DETECTION AND MONITORING BASED ON FUSION OF EARTH OBSERVATION AND INSPIRE DATA	473
<i>Sergio Samarelli, Vincenzo Massimi, Luigi Agrimano, Daniela Drimaco, Raffaele Nutricato, Davide Oscar Nitti and Maria Teresa Chiaradia</i>	
SAR ALTIMETRY PROCESSING ON DEMAND SERVICE FOR CRYOSAT-2 AND SENTINEL-3 AT ESA G-POD	477
<i>Jérôme Benveniste, Salvatore Dinardo, Giovanni Sabatino, Marco Restano and Américo Ambrózio</i>	
BROADVIEW RADAR ALTIMETRY TOOLBOX	481
<i>Albert Garcia-Mondejar, Roger Escolà, Gorka Moyano, Mònica Roca, Miguel Terra-Homem, Ana Friaças, Fernando Martinho, Ernst Schrama, Marc Naeije, Américo Ambrózio, Marco Restano and Jérôme Benveniste</i>	
ATOS CODEX SPARKINDATA: PROMOTING USER UPTAKE OF AN EARTH OBSERVATION APPLICATION MARKETPLACE	483
<i>Joanna Emery, Philippe Lattes, Chadi Jaber and Kyriakos Konstantopedos</i>	

INTERNET MAJOR ACTORS TECHNOLOGIES APPLIED TO METEOROLOGICAL SATELLITE: STREAM DATA APPLIED TO DATA PROCESSING

Alain Montmory (1), Laure Chaumat (1)

(1) Thales, 290 Allée du Lac – 31670 Labège - France

ABSTRACT

Big Data technologies have been at the heart of Thales strategy for already six years. In French South West division, development teams are mainly working with CNES and EUMETSAT actors and have acquired business knowledge on space ground segments. Thales gets its first real experiences mixing Big Data and space mission by cooperating with the CNES on GAIA [1] ESA mission. Together we were the firsts to use the famous Hadoop technology on space data to realize the CNES Data Processing Center. Then we applied our knowledge through several projects for EUMETSAT: L2PF processing platform for geostationary satellite and PDAP which deals with a polar satellite – both of these projects are still in progress and are not in operational mode as GAIA. The developed Data Processing Infrastructure (DPI) uses the stream technologies set up by Twitter (Storm [2]) and LinkedIn (Kafka [3]) and a micro services approach builds onto the Docker[4] Ecosystem. This paper will present the use case where we deploy these technologies and the feedback on it.

Index Terms — STORM, KAFKA, ELK, DOCKER, CNES, EUMETSAT, L2PF, PDAP

1. MISSIONS OBJECTIVES

The EUMETSAT's mission is to establish, maintain and exploit European operational meteorological satellite systems. A further objective is to contribute to operational climate monitoring and detection of global climatic changes. To fulfill these objectives, two types of meteorological satellites are needed: Geostationary orbit satellites, vital for forecasts up to a few hours; Polar orbit satellites, critical for forecasts up to 10 days.

The MTG (Meteosat Third Generation) and EPS-SG (Eumetsat Polar System –Second Generation) systems will provide Europe's National Meteorological Services and the International Users and Science Community, with an improved imaging and new infrared sounding capabilities for both meteorological and climate applications. The objective of these systems is to provide continuous high spatial and temporal resolution observations and geophysical parameters of the Earth System derived from direct measurements of its emitted and reflected radiation using satellite based sensors. Information based on imagery of the globe and hyperspectral sounding of the atmosphere will also deliver unprecedented volumes of data.

2. WHY STREAM TECHNOLOGIES

The MTG system is a constellation of three MTG satellites; which have a continuous downlink throughput of about 600 Mbps. The Level-2 Processing Facility (L2PF) is part of MTG ground segment. L2PF is fed by level 1 system IDPF-I/IDPF-S data flow on an H24 basis. This continuous flow contains data of 4 instruments which are the Flexible Combined Imager (FCI), the Lightning Imager (LI), the Infra-Red Sounder (IRS) and the Copernicus Sentinel 4 Ultraviolet Visible Near-infrared (UVN). Due to the constant and high data rate between the IDPF-I/IDPF-S (level 1) and L2PF (level 2) the processing can only be handled by stream technique (as opposed to batch technic used on GAIA mission).

3. THE DPI, HOW TO EMBED SCIENTIFIC PROCESSING FOR 20 YEARS MISSION LIFETIME

Lessons learned from GAIA led to a major design decision: to separate data from processing (i.e. it is the Processing Infrastructure (PI) which brings the data to the processing code which embeds the scientific algorithm). Such design allows switching, if needed, the underlying technologies without changing the scientific code. Currently, the streaming engine is based on Storm. In the future, it could be changed if Storm becomes obsolete or if there is a need to introduce a batch capability (example Flink framework which allows Batch and Stream capabilities). The whole DPI design relies on very simple interfaces called IChannel and ITuple which allows abstracting the infrastructure (IChannel) and the data (ITuple):

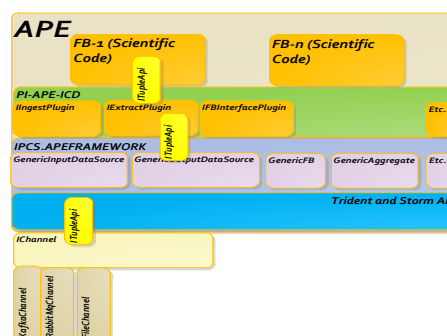


Figure 1: DPI Layered design

The scientific code (APE: Algorithm Processing Element) is embedded into plugins which are called by APE framework.

4. A GRAPHICAL TOOL TO HELP SCIENTIFIC CODE DEVELOPMENT

Comparing to other systems which deal with file in input/output, the DPI implies a new paradigm for parallelism based on input data segmentation. To help the scientific developer, a graphical tool is provided which allows to describe graphically the APE by a chain of data operators similar to those which exist in the SQL language (groupBy, Join, forEach etc...). These operators are made available to the scientific developer by an abstraction layer called the “APE framework”. These operators exist in quite almost all Big data framework (Hadoop (Cascading, Pig), Spark, Flink). The “APE framework” hides the underlying runtime engine; there is no link at all between the scientific code and the runtime execution engine (Storm or other) which allows the replacement of this runtime engine if needed during the 20 year mission lifetime.

The graphical tool generates the APE code for both Standalone and Production environments. It allows the APE developer to test their APE in a mode accessible on a single laptop using the same interfaces as those which are used in production.

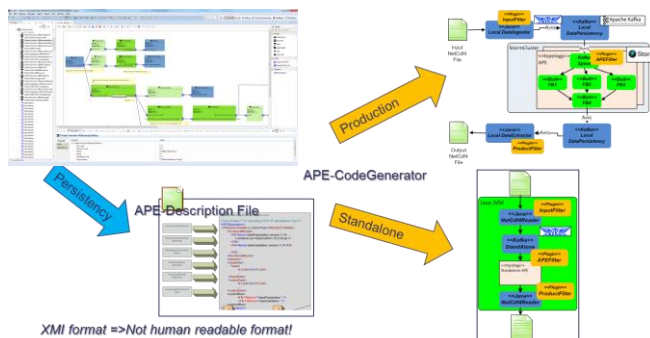


Figure 2: Both Standalone and Production generation mode

5. PACKAGING, DEPLOYMENT: THE MICRO-SERVICES ARCHITECTURE

As the data could come from a geostationary satellite (MTG mission), the data flow is continuous, so an availability of 99,8% is requested for the L2PF. In this context, the packaging, deployment and switch version process shall be as smooth as possible (the deployment of a new version shall have no impact on the running processing and the version switch shall be as fast as possible (few seconds)). To deal with these constraints the DPI is built upon a solution based on micro services architecture:

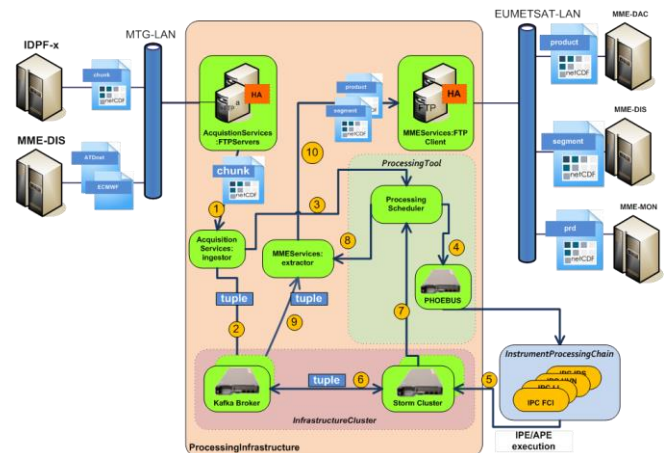


Figure 3: DPI micro-services architecture

Each DPI component plays a single and simple role (the micro service) and is packaged inside a Docker Execution Container (DEC). All the DEC are assembled using Docker Compose tools and deployed onto a Docker cluster based on Docker Swarm. Thanks to the Docker Compose tool, each component is scalable in 2 or more instances for performance and availability reason (2 is the minimum). This flexible deployment allows to cope with different processing speed from 1 for Near Real Time, 5 time faster for reprocessing and 30 expected for the Algorithm Fast Processing (a tool used in development scope to test the APE on a “production like” cluster).

The DPI components are deployed in a private network brought by Swarm overlay network drivers. Some components have also an externally reachable network address (example input FTP servers) setup by pipework script. Docker ecosystem is also used for the build, deploy and run process. The build result is a set of Docker images which are pushed into the Operational Configuration Management Registry. Then these images are deployed onto the “live” environment following the “descriptor.yaml” file which describes all the images used on the target “live” environment. There are four possible “live” environments: OPE, VAL, IVV, REP (Reprocessing). The Docker ecosystem is also used to package “long term configuration data” like Digital Elevation Model which are embedded into versioned Docker Data Container (DDC). The DDC are deployed as other Docker image and then mounted into DEC to give access to the data (the DEM file in our example).

6. STREAM PROCESSING CONCEPT

The stream processing concept is based on the concepts of key and partition. In scientific processing it is frequent that a data needs its neighborhood to be processed. This grouping is assured by allowing a Key which represents a scientific criteria, for example a Swat or a geographical zone id. The partition allows to dispatch the data onto several

computers..The figure below shows a typical processing flow.

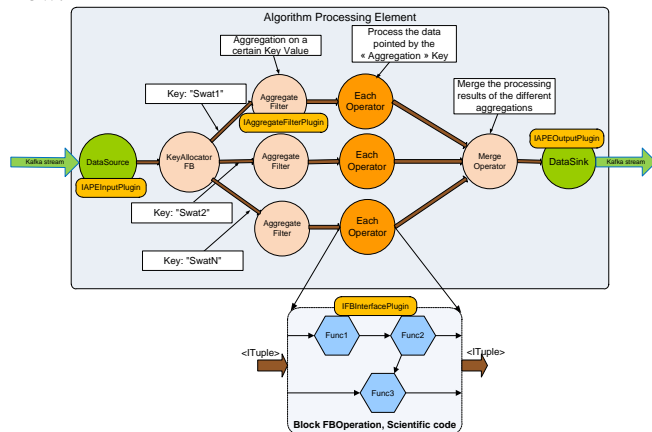


Figure 4: Typical APE processing flow structure

The data are injected from Kafka stream into the APE using a “DataSource” operator, then the data are partitioned by key using a “keyAllocator” FunctionalBlock. The data are aggregated using an “AggregateFilter” operator. When an “Aggregation” is filled (the filter answered yes “the aggregation is full”, the aggregate data are processed by the “Each” operator, which embeds the real “scientific processing code”. Each of the operators listed above embeds “business” code through “plugin” interface. The plugin code is provided by the scientific team. This scientific code is independent of the underlying “runtime processing framework (Apache Storm in case of L2PF/PDAP), because the abstraction of “logical” operator like “DataSource”, “DataSink”, “AggregateFilter”, “Join”, “Merge” etc.. hides the “runtime processing framework”. This layer of “logical operator” is called the “ApeFramework”. This layer authorizes (if needed) the replacement of the “runtime processing framework” by another one (for example, to prove the design concept, an ApeFramework port has been done on Apache Flink), which is important to ensure the 20 years mission lifetime on this kind of mission.

Another key concept of the DPI design is the data normalization around the concept of ITuple. A ITuple is a tree of {key|value} where value could be another ITuple. The figure 5 : Structure of an ITuple below synthesizes the concept.

All the data (L1 data chunk, AuxiliaryData etc..) injected into the DPI are normalized into ITuple, which allows a simple definition of the interfaces at source code level (input and output are ITuple). The ITuple is well adapted to the NetCDF input data which are also a tree structure (but the NetCDF format is not the only one used by DPI, there are also GRIB and BUFR formats, which implies the need of normalization.

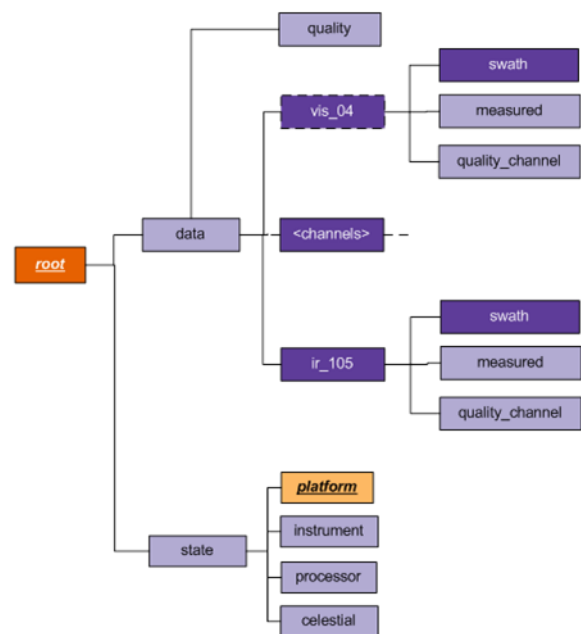


Figure 5: Structure of an ITuple

7. DATA MANAGEMENT CONCEPT

The L2PF DPI data management concept is based on Kafka for data distribution and on Elastic search Logstash Kibana (the ELK [5] framework) for data indexation.

The Kafka system allows to store and distribute the live data in “publish/subscribe model” but at the difference with other messaging system like AMQP (RabbitMq, ActiveMq) it is not a server which pushes the data to the client but it is the clients which retrieve the data at their own rate, the current retrieved record offset is stored reliably, for each client, into a cluster shared context (Apache Zookeeper for DPI). This design optimizes the performance of the whole system because the kafka broker does not have to deal with slow clients.

The data retention is managed independently of the data consumption and on per Topic (a Topic is a logical address on which data are pushed and retrieved) basis. For example data can be kept locally for 24 hours to take into account the possible unavailability of the EUMETSAT long term archive system (The DAC : Data Archive).

Kafka is an efficient data distribution system but it does not handle data selection filter, it is mainly a FIFO model.

To handle more complex data selection scheme with respect to the scalability of the whole system, the data are also indexed in a Elastic Search cluster. It encompasses the L1 Input chunk data, the LOG generated by the DPI components and the APE, the processing parameter issued by the APE and infrastructure parameter gathered by SNMP agent on each docker hypervisor. It allows to display LOG report and processing parameter significant values output by

the running APE. The LOG and Processing Parameters display allows an operational monitoring of the L2PF facility through dedicated MMI operates by Controller Operator in EUMETSAT premise. The following screenshot shows the LOG display and a Parameter curve displays by Kibana view embedded in the Eclipse RCP MMI application:

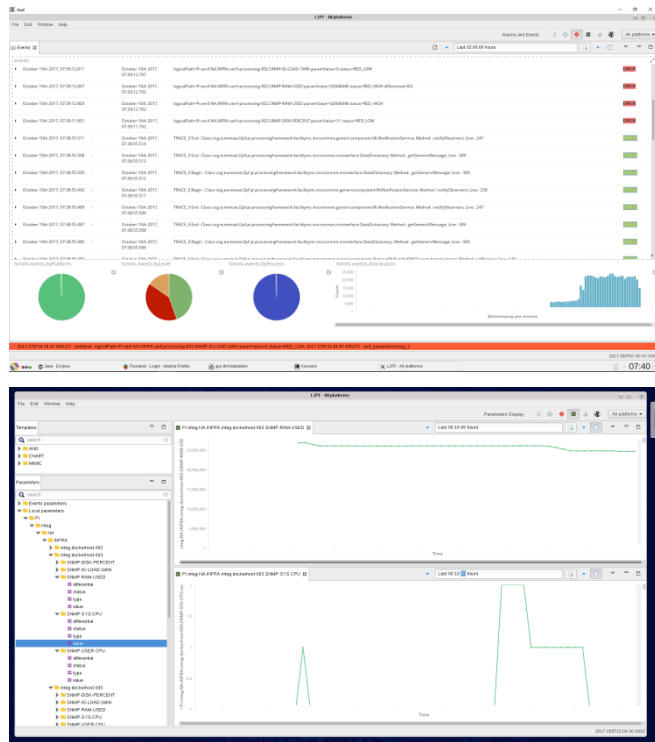


Figure 6: Indexed data display through Kibana

8. HIGH AVAILABILITY CONCEPT

The DPI is 24H 7/7 system with a requested availability of 99.8%. These constraints have driven the Open Source components selection:

- ✓ Kafka has a built-in data resiliency with a transparent data replication (which also improves the read speed),
- ✓ Storm has the concept of Directed Acyclic Graph, each input record ingestion consequences are monitored and acknowledged at record level, If a record is not acknowledged the Storm supervisor relaunches the record processing (which may occur on different computers). Of course the reprocessing does not occur indefinitely (3 tentatives) to deal with malformed data.
- ✓ The Docker ecosystem, with the Swarm clustering mode offers also a convenient way to ensure processing resiliency. If a Docker Container fails it is relaunch by the Swarm. If an entire Docker Host is dismissed (for example for maintenance

activities), the whole Docker Container running on it are moved on another Docker Hosts (the hardware design take it into account by providing 33% free docker slot. The following schema synthesizes the Docker Container migration:



Figure 7: Docker Resiliency brings by Swarm

With these concepts, the result is quite similar to a VMWare ecosystem without the penalty performance of the Virtual Machine hypervisor (measured at 15%) and the extra cost of VMWare tool licences (VSphere, CloudDirector etc..).

9. CONCLUSION

The technologies used by DPI, although they are not intended to be used in the context of scientific processing at their origin have proven their performance, scalability and resilience features. The Docker ecosystem, despite its youth demonstrates that the slogan “build once, runs everywhere” is not only a slogan but also the reality.

10. REFERENCE

- [1] “.Data Management at Gaia Data Processing Centers”
- [2] “Apache Storm” <http://storm.apache.org/>
- [3] “Apache kafka” <https://kafka.apache.org/>
- [4] ”Docker - Build, Ship, and Run Any App, Anywhere” <https://www.docker.com/>
- [5] “The Open Source Elastic Stack” <https://www.elastic.co/products>

FROM HADOOP MAP/REDUCE TO SPARK: GAIA AND IMAGE PROCESSING USE CASES

Guillaume Eynard-Bontemps², Olivier Melet², Hugo Palacin¹, Florent Ventimiglia¹, Louis Noval¹

¹Thales, 290 allée du Lac, 31670 Labège

²CNES, 18 Avenue Edouard Belin, 31400 Toulouse

ABSTRACT

The Gaia DPCC uses a Hadoop/Cascading based framework in order to compute scientific data since 2011. This solution is fully functional whereas a breakthrough in Hadoop ecosystem called Spark introduces a theoretical improvement factor of up to 100. This new system manipulates data through high level API and optimizes in memory computation instead of historical data exchanges through hard disk drives. This paper will present some results and lessons learned from using Spark though Gaia data manipulation queries and another use case of image processing.

First, we will introduce Gaia data and processing context followed by a brief overview of Spark. Then, we will present our implementation choices and compare performances between Cascading and Spark when running typical Gaia workloads. We will then talk about a Spark use case on Image processing. We will finally conclude and see what our next objectives are.

Index Terms— Hadoop, HDFS, Cascading, RDD, Spark, YaRN, Ground segment, Image processing, Gaia

1. GAIA

The Gaia mission aims at building an extremely precise three dimensional map of our galaxy. CNES Data Processing Center handles hundreds of billions of objects of different types into kinds of big tables or collections. Each record size is between 100B to several hundred KB. This makes Gaia data manipulation very suitable for Big Data frameworks such as Hadoop, hence the initial choice.

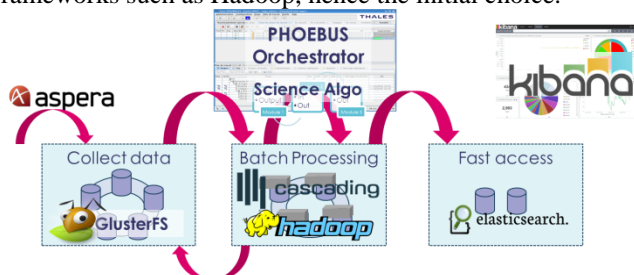


FIGURE 1: GAIA DPC GLOBAL ARCHITECTURE

Gaia processing chains need to perform really complex queries on several input tables. These queries are often quite comparable to what can be done using SQL language.

Implementing those queries in pure Map Reduce is nearly impossible. That's why Cascading framework, which can be compared to Pig or Hive, was chosen to abstract the complex data manipulation layer (join, filter, ...).

Hadoop and cascading are the core of Gaia CNES Data Processing Center, specifically the System of Accommodation for Gaia Algorithm (SAGA) that effectively processes the data (). On a higher level is Phoebus, an orchestration tool that allows chaining Cascading pipes through "steps". Each pipe is effectively executed as one or more Map/Reduces jobs. Scientific modules can be encapsulated in either a Map or Reduce phase, as you can see on Figure 2

It is also important to highlight the specificities of a Hadoop cluster (commodity hardware, meaning low costs), and also the fact that DPCC deployment is entirely automatized using DevOps tooling (Puppet and Foreman).

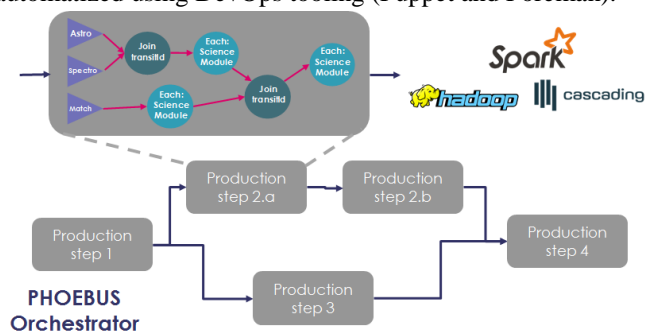


FIGURE 2: SCIENTIFIC MODULES, CASCADING AND WORKFLOWS

2. SPARK

2.1. Spark APIs

Spark is a processing framework which is replacing Map Reduce processing layer from Hadoop. It provides a higher level and simpler base API layer, a complete integrated libraries stack (SQL, Streaming, Machine Learning ...), and improved performances with its optimized engine and by working in memory.

Since Spark 2.0, there are two main programming APIs for data manipulation: Spark core with RDDs, which are basically a collection of objects or records, distributed across all computing nodes; Spark SQL with Datasets,

usually faster and more convenient [1], with a more structured way of manipulating data (SQL like).

2.2. Spark vs Hadoop/Cascading vs the rest of the world

Spark has its own engine and do not rely on Map Reduce like Cascading. It manages the flow execution itself, recursively decomposing it into multiple applications, stages and tasks. The execution engine natively handles acyclic graphs (DAG), and heavily relies on in memory computation which greatly optimizes data transfers. On the contrary, a Cascading Flow is automatically split into one or more Hadoop Map Reduce jobs at execution time, and relies on disk writes for data exchange between each of these steps.

It is also worth noting other frameworks which are concurrent of Spark. Flink is primarily a stream processing framework, but it provides also strong batch processing APIs. It is really close to Spark with all its optional libraries (Batch, Stream, Machine Learning, SQL, Graph ...). It is a powerful tool which could be studied too, especially because of its Cascading compatibility. But at the time of the study, Spark was much more improved and thus, a more reliable choice. Dask is a python framework which handles big collections like Spark, but also complex tasks scheduling. It has not exactly the same purpose and was not well suited for Gaia needs, but it could be adapted for image processing.

2.3. Programming Languages

One of the advantages of Spark is that it is possible to develop applications using several programming languages. Indeed, it is possible to use at its convenience Java, Scala or Python. It is noteworthy that Spark is natively coded in Scala, so it is more efficient to use Scala than another language, especially compared to Python which can lower performance (up to a factor of ten in some cases).

Because Gaia DPAC relies on Java 1.8 environment only Scala and Java8 API are relevant for us. Both pros and cons are very close and we decided to implement case studies on Scala over Java8 because of greater documentation available online.

3. SPARK ON GAIA

3.1. RDD or DataSet?

We already exposed advantages of DataSets over RDD. Nevertheless, their use seems to be compromised because of GaiaRoot type itself. Object hierarchy in Gaia data model is complex and some attributes are not primitive types. This is not natively supported by DataSets: there would be a need to implement a dedicated wrapper or to adapt each GaiaRoot sub type. DataSets have been developed for structured data (with a known schema) such as databases, JSON files, or structured text tables. RDD will thus be used for our tests.

3.2. A word on file formats, object representation and serialization

The inability to use Dataset also highlights a point linked to storage format efficiency. Gaia objects storage in Hadoop is based on the old SequenceFile format, with Java Serialization mechanism applied on GaiaRoot objects. With improved serialization / deserialization methods defined or better with a more structured schema for Gaia tables, storage efficiency and processing time could be greatly improved. Some tests have been performed by just switching to Kryo Serialization, but it did not improved performance a lot. This point should be studied further given some results obtained by other institutions [6].

3.3. Pipelines on Cascading and Spark

Cascading and Spark both represent data manipulations as graphs. It should be noted that the high level views of generated graphs are quite close, concepts are the same: Pipe and RDD, Operations and transformations. Spark RDD API seems a little less verbose and simpler to use and to understand, but it has also one drawback compared to Cascading Pipe API: it is based on Key Value mapping when performing joins or grouping. This means that every time we need to do one of these operations, we need to write some boilerplate code to extract the appropriate Key from the records of our manipulated collection. With Cascading, or Dataset API, this limitation is not present.

3.4. Study use case and environment

We compared performances of integrated Cascading/Hadoop subassembly over the same operations through RDD Spark API. In both case we removed all scientific modules calls. The next diagram illustrates the graph of operations that was applied on data.

Tests have been performed on Gaia integration platform; resources were limited to 20 cores and 80 GB of RAM on 4 calculation servers (about 50% of the platform resources). Because of concurrent uses of the platform for project purpose, we used 2 indicators to measure job execution time: an external clock (user time) and the total amount of CPU time spent by each thread. This method prevents from having inconsistent results due to parallel resource uses.

We also used several datasets of different input size: from 150 MB of compressed data to 240GB. We observed a compression factor varying between 2 or 10 depending on the object type, but a reasonable mean would be 3. So there were between 450MB and 720GB of uncompressed input data.

We studied two main queries; CU4 SSO_LT ingestion subassembly and CU6 CI_PP subassembly (See Figure 3: CU6 CI_PP subassembly, which is the more complex.

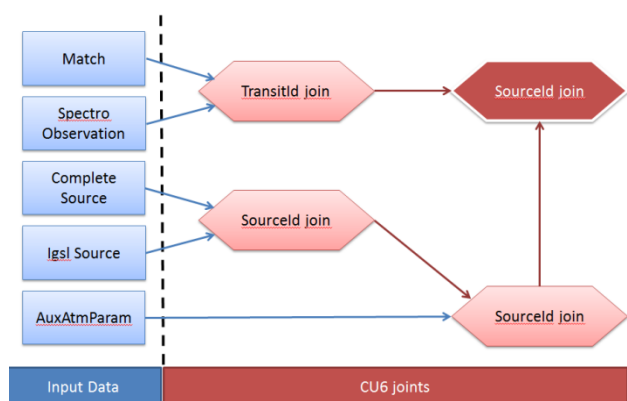


FIGURE 3: CU6 CI_PP SUBASSEMBLY

3.5. Performance results

We noticed a clear gain with Spark for datasets that fit in memory, but we saw a decrease in Spark's benefits compared to Cascading as the volume of data to process grows. This can be explained by the fact that Spark works in memory and that when it is full it must spill to disk which penalizes performances. Hadoop also has a big drawback: it has to start JVMs for each task, which heavily penalizes it on small datasets, whereas Spark uses executors started once and for all the job duration.

For the biggest dataset which needed about ten times more storage capacity than the memory available, Spark advantage is really small in term of CPU usage. Measuring the overall user time, we see a bigger advantage for Spark: we noticed that we still have a 36% gain in execution time for Spark (see **Erreur ! Source du renvoi introuvable.**). It shows that Spark really optimizes the use of available CPUs. Moreover, the DPCC Operational platform has 19 TB of RAM allocated to the compute nodes, enough for a lot of use cases.

3.6. Spark operational use

Spark has proven very easy to deploy and to use on an already existing Hadoop cluster. Deployment is just a matter of installing the application all nodes. Moreover, it is very well integrated into YaRN. It means that it is also quite straightforward to use the same cluster to submit Cascading/Hadoop or Spark applications at the same time. Last but not least, Spark provides a really good web UI that allows monitoring and analyzing Spark applications. The fact that Spark sees a complete dataflow as only one application is really a benefit for operators compared to Cascading and its decomposition of flows in several MapReduce jobs.

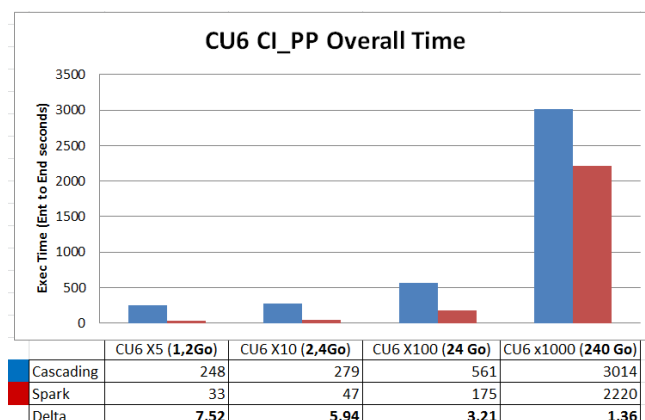


FIGURE 4: CU6 CI_PP USER TIME

3.7. Next steps

Gaia Spark study has been split in two main parts. The first one is completed and the results are presented here. Even if performance results are promising, they have to be tempered by end-to-end chain performances: Data insertion/extraction and scientific computations are not faster and we cannot afford to re-design existing chains for now.

The last part of the study, still in progress as we are writing this paper, aims to implement a prototype in order to let scientific algorithm integration team work with Spark for Gaia next generation chains composed of iterative processing. Data iterations and Map/Reduce are not close friends, whereas Spark is really good at it when dataset fits in memory. We are also performing a more complete analysis of Gaia chains in order to detect if Spark could improve some algorithmic parts. It could be the case for Cascading flows generating a lot of MapReduce jobs. Finally, aside from performance concerns, Spark would provide a good way to execute Machine Learning algorithms, for which classical MapReduce is not well suited.

4. IMAGE PROCESSING WITH SPARK

4.1. Context

We experienced Spark in another context: earth observation image processing. Most of CNES processing chains are currently executed on an HPC cluster, with shared centralized storage and computing resources. Complex workflows are designed as a chain of independent jobs. All the data synchronization between jobs is done on an IBM Spectrum scale system. This mechanism induces a big overload of IO on the storage space, with sometimes not optimized algorithms. This is especially true in a context of always increasing image resolution and number, leading to an exploding volume to handle. Spark (among other technologies) has been tested in order to benefit from its horizontal scalability and data manipulation APIs. We made

use of its ability to design complex graphs for distributing tasks which allows taking advantage of distributed memory and computing node local storage, avoiding the need of data synchronization.

4.2. Use case experiment

An image processing chain can be represented as a graph of process, which are either sequential or can be treated in parallel. Between them are some synchronization points. We defined a typical chain (as depicted in Figure 5: Implemented image processing chain), composed of standard processes (decompression, ortho rectification ...).

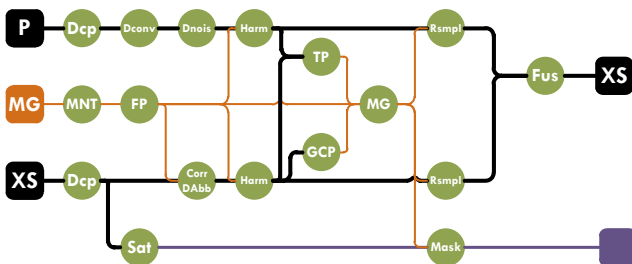


FIGURE 5: IMPLEMENTED IMAGE PROCESSING CHAIN

Three main points have been tested:

First, the implementation of each given algorithm using Spark. This required the use of RDD containing images tiles as items. In order to implement algorithms that needed a tile and its neighboring zone, some data manipulation involving reductions and aggregations have been used. An ImageRDD abstraction has been defined to handle split large images as a unique object.

Secondly, for CPU time intensive algorithm having legacy libraries written in C++, we defined a standard interface to be able to call them from Java. This really eased legacy code integration and we observed a notable improvement of performances.

Finally, thanks to the ImageRDD abstraction, we were able to chain several algorithms, using Spark lazy evaluation capability to model our workflows as Direct Acyclic Graphs.

Some key takeaways from this experiment: Spark really eases the images tiles manipulation with its abstraction; it handles high level scheduling; scheduling and data moving between nodes is only a tiny part of the computation time.

5. CONCLUSION

This study confirms on both Gaia data and image processing use cases that Spark is a new reference in term of Big Data processing.

It shows really good results on Gaia when working in memory (between 3 and 8 times better for reasonably sized problems), and still has a 30% advantage when working on disk. But it is not as high as expected considering marketing

promises [3]. It is also easier to program, with a really good documentation and high level API, despite the limitation on Key Value manipulation with RDD. It has also strong monitoring tools. Compatibility with Hadoop environment makes it a good candidate to develop new processing algorithms for Gaia. The still on-going study will help to identify which use cases are the most suitable to take advantage of Spark according to different criteria: iterative processing, data volume, memory consumption...

On Image processing, it has proven to be really helpful both in terms of data distribution and in terms of algorithms chaining and scheduling. Performance results are very promising and it should really help for having a scientific chain compatible with both on premises cluster and public cloud infrastructures.

6. REFERENCES

- [1] T. White, "Hadoop, the definitive guide", 2 nd ed, Oreilly, 2010.
- [2] <http://www.kdnuggets.com/2016/02/apache-spark-rdd-dataframe-dataset.html>
- [3] <https://databricks.com/blog/2016/01/04/introducing-apache-spark-datasets.html>
- [4] <https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html>
- [5] Michael Hausenblas and Nathan Bijnens, inspired by Nathan Marz, <http://lambda-architecture.net>, 2015.
- [6] Zbigniew Baranowski, Performance comparison of different file formats and storage engines in the Hadoop ecosystem, CERN, [Jan 27, 2017](http://arxiv.org/abs/1701.02501)

BIG DATA PROCESSING USING THE EODC PLATFORM

S. Elefante¹, V. Naeimi¹, S. Cao¹, I. Ali¹, T.S. Le¹, W. Wagner^{1,2}, C. Briese²

¹ Department of Geodesy and Geoinformation, Vienna University of Technology, Vienna, Austria

² EODC Earth Observation Data Centre for Water Resources Monitoring, c/o University of Technology, Vienna, Austria

ABSTRACT

The goal of this work is to optimize the processing of large earth observation (EO) data sets on the Earth Observation Data Centre for Water Resources Monitoring GmbH (EODC) platform, which is a collaborative cloud infrastructure for archiving, processing, and distributing EO data. EODC provides a comprehensive environment through which any worldwide user can remotely (*i*) access EO data, (*ii*) process them in a dedicated workspace, and (*iii*) visualize the final products. Different tests involving the processing of the whole ENVISAT ASAR GM archive, and parts of the ASAR WS and Sentinel-1 archive were performed. The experiments showed the computational capability and flexibility of “big data” processing using EODC’s functional design of the different IT components (storage, cloud platform, supercomputing).

Index Terms— *Big data, parallel computing, SAR, high performance computing (HPC), soil moisture*

1. INTRODUCTION

The availability of large amounts of Synthetic Aperture Radar (SAR) data stemming from the completed ENVISAT satellite mission and the Sentinel-1 satellites constellation represents a major opportunity for the worldwide scientific community. Earth observation satellite data are being widely investigated for many operational and scientific applications. One of the preeminent applications of Sentinel-1 over land is to monitor surface soil moisture that is considered as a key parameter for flood forecast and numerical weather prediction systems [1] [2]. However, the processing of a Tera- to Petabyte-scale SAR dataset can only be achieved by following the well-established paradigm of distributed computing. Thus, the processing of the whole ENVISAT archive and of the Sentinel-1 data in a time frame that allows researchers to make scientific conclusions within a reasonable time span must be carried out in a computing environment that fully supports parallel processing. The high performance computing (HPC) system represented by the Vienna Scientific Cluster 3 (VSC-3) [3] suits this aim well. The VSC-3 is a large distributed platform installed in summer 2014 in Vienna. Due to the massive volume of the Sentinel-1, accessing and downloading the raw data are not trivial issues.

To effectively deal with these challenges, the paradigm is to bring the processing algorithms close to the data as well as to the computing resources. EODC [4] provides a feasible solution to this paradigm. EODC is a private-public partnership aiming to deliver a collaborative cloud infrastructure for archiving, processing, and distributing earth observation (EO) data. Through multi-national partnerships from science, public and private sectors, users can get direct access to Sentinel data storage and run data-intensive geoscientific models. One of the services offered by EODC is the capability of performing data processing through VSC-3. The EODC environment is, therefore, a comprehensive infrastructure in which users have the possibility to directly access and process EO data with their own algorithms and retrieve the final products.

The aim of this work is to assess the potential of processing big data for remote sensing applications with an attention to the TU Wien soil moisture retrieval processing chain. The results of three case studies, which were performed by using the EODC infrastructure and the VSC-3 high performance computing platform, are discussed.

2. DATA ACCESS

Employing VSC-3 in the EODC environment gives users access to two different kinds of storage volumes: (1) storage provided by the VSC-3 that can be used for intermediary (temporary) storing files during processing, (2) the dedicated EODC storage for accessing EO data in the public part of the archive and for storing results in private folders. The technical features of the two volumes are described below:

- (1), the VSC-3 distributed volume runs on the parallel clustered file system BeeGFS [5], which is installed on 360 spinning disks connected through 160 Gb/sec bandwidth.
- (2), the EODC storage is based on IBM Elastic Storage Server (ESS) [6]. It features the high-performance GPFS clustered file system with a current net disk capacity of 2 Petabytes (PB).

The VSC-3 cluster and the EODC archive are connected through an InfiniBand fabric (8×40GB/s bandwidth). Additionally, an IBM TS4500 Library is used to store less frequently accessed EO data and for backups. At the moment, the library consists of 6 Drives including 4 PB tapes.

The EODC public library currently features more than 2.000.000 individual Sentinel-1, -2 and -3 data sets.

3. PROCESSING AND VISUALIZATION

The main advantages of the EODC environment are the possibility to remotely access one's workspace from anywhere and at the same time directly access all available EO data without the need to download them to a processing system. Users may connect from their own PC to

(1) the EODC private cloud infrastructure, which provides users with virtual machines (VM) for developing and testing the methods, as well as visualizing the results on small spatio-temporal extents (Fig. 1). The cloud environment comprises 200 physical cores, 5,3225 TB RAM and 122 TB disk storage.

(2) the VSC-3, through dedicated login nodes, and, subsequently, the compute nodes. The VSC-3 is an advanced HPC system and consists of 2020 nodes, each equipped with 2 processors (Intel® Xeon® Processor E5-2650 v2 from the Ivy Bridge-EP family). Totally, there are 32.000 physical cores internally connected with an Intel QDR-80 dual-link high-speed InfiniBand fabric. The CentOS Linux release 7.2.1511 is installed on each node as operative system. The Simple Linux Utility for Resource Management (SLURM), which is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters, is installed as middleware. SLURM organizes the access to the computing nodes through the management of a queue of pending work and provides a framework for executing and monitoring jobs. The VSC-3/EODC infrastructure is shown in Figure 1.

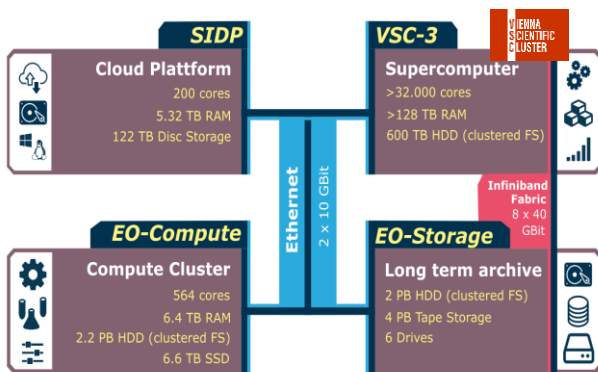


Figure 1. EODC infrastructure

4. PROCESSING RESULTS

The big data processing was performed using VSC-3. For comparison, three different Level-1 SAR datasets including data acquisitions from ASAR GM, ASAR WS and Sentinel-1 have been processed on the EODC environment using the SAR Geophysical Retrieval Toolbox (SGRT) [7] developed at the Vienna University of Technology. SGRT produces a

series of different geophysical parameters through time series analysis of satellite SAR data. SGRT is written in Python programming language and includes external software modules, e.g. ESA's Sentinel-1 toolbox (S1TBX) for SAR data geocoding, radiometric corrections and calibration. The full archive of the ENVISAT ASAR GM data (size of about 1.6 TB) and around the 76.7% of ASAR WS archive, 86,199 files with a total size of about 15.7 TB were successfully pre-processed (geocoded and georeferenced) during case-1 and case-2 (Table 1).

Table 1. Results of the three case studies.

	case-1	case-2	case-3
SAR data product mode	ASAR GM	ASAR WS	S1A/B GRD
Spatial resolution	1 km	150 m	10 m
Number of images for job / Total Number of jobs	8 / 23,703	2 / 43426	1 / 63,657
Total input data files size	≈ 1.6 TB	≈ 15.7 TB	≈ 73.6 TB
Total output data files size	≈ 2.1 TB	≈ 60 TB	≈ 129 TB
Percentage of the archive that has been processed	≈ 99 %	≈ 76.7 %	≈ 20% (October 2014 - July 2017)
Averaged processing time (seconds/MB)	9.18	2.39	2.69
Aggregated elapsed time including SLURM queueing	≈ 3.5 days	≈ 15 days	≈ 22 days
Theoretical estimated elapsed time using only 1 node	≈ 167 days	> 1 year	> 1 year

Part of the Sentinel-1 image stack has also been pre-processed, almost 20 % of the all images acquired between October 2014 to July 2017 (case-3 of Table 1). All data were hosted on the EODC archive. Table 1 shows the detailed results of the processing including respective run-time estimates and measurements. S1 A/B data processing has been performed processing each time only the new incoming images, the aggregated time is referring to a rough estimation of the total time considering all these different runs. The processing was conducted by consecutively transferring data from the EODC archive to the RAM disk created on the VSC-3 computing nodes, performing the necessary computation steps and saving the results either *i*) on VSC-3 storage and then moving data back to EODC archive, *ii*) or directly to the EODC volume. Our tests showed that both are almost equivalent in terms of computational performance as the VSC-3 internal storage and the EODC volume are very similar concerning I/O access. Indeed, the writing rate towards the VSC-3 BeeGFS distributed volume is around 1.2 GB/sec and 450 MB/sec while for the IBM ESS storage is about 1.4 GB/sec and 500 MB/sec for respectively simultaneously running 1 and 5 jobs (See Table 2). However,

these values have been experimentally evaluated computing the average between the processing time that have been computed running the test many times; the infrastructure is employed by many users and a clear benchmark evaluation is not practically feasible.

In all the three cases as shown in Table 1, the large number of files (for instance more than a million in the case of ASAR GM when also intermediate files and results are considered) makes it burdensome to save the output directly in the storage distributed volume (both VSC-3 and EODC). The storage devices were not capable to efficiently handle the multiple I/O operations that were simultaneously requested from all the computing nodes. All attempts for storing the output data directly on distributed hard disk failed due to fatal crashes of the nodes. This also occurs because distributed storage volumes are designed to have an optimal performance for files with the size of around 1-2 GB, and the usage of small files increase the chances of crashing. Thus, the images were divided into small groups of 8 and 2 images per job for ASAR GM and WS, respectively, only 1 image for S1, and copied in the RAM disk of the selected VSC-3 computing node. Also, the output was temporarily cached in the RAM disk of the node and only at the end of the processing the output was copied on the hard disk (persistent BeeGFS or ESS storage volumes). This computational strategy has minimized the number of I/O operations which result in reducing the stress for the storage system. The maximum number of images (i.e., 8 ASAR GM, 2 ASAR WS and 1 S1A/B) that a single node can process was constraint based on the available RAM on each node.

By exploiting the I/O and CPU processing times, the maximum number of simultaneously running nodes can be easily determined in order to fully utilize the available bandwidth without risking to saturate the network and disk. As shown in Table 2 the writing/reading rate for the VSC-3 and EODC storage is about 1.2 and 1.4 GB/sec respectively. This means that the time required to read and copy a S1A image (of size around 2 GB) on the RAM disk of the computing node is approximately 2 seconds. This occurs only if no other nodes are performing I/O operations (Table 2 shows that when 5 CPUs are simultaneously writing, the I/O rate drops to 500 MB/sec). The output (4 GB) of the processing chain is doubled the size of the input (2 GB), which will double the time to transfer and write this processed data from the node RAM disk to the storage.

To fully utilize the available resources in an optimal way and avoid the fatal breakdown of network and storage, a new strategy of calculating the maximum number of nodes that can be ran in a sustainable manner, is defined. Where, when one node is finished with reading or writing task the another node can immediately follow to start reading or writing.

Figure 2 shows a graphical abstract of the proposed strategy; when node 1 has finished to read data (block horizontal dashed), node 2 can start to perform its data reading and so on for node 3 until node n. Same evaluation is also valid for

the writing operation, and therefore the ratio between CPU processing time and the total I/O time provides the number of nodes such as when a node has finished its own I/O operations another can start without I/O overlapping.

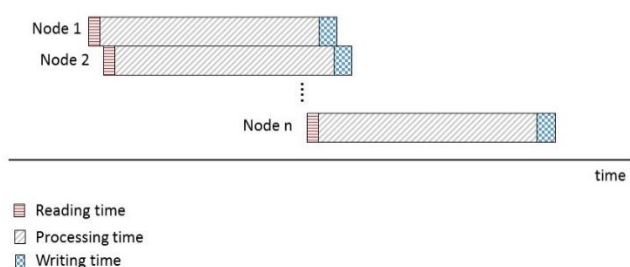


Figure 2. Logic to evaluate the optimal number of nodes

Taking into account that the S1 single-scene average pre-processing time is around 40 minutes, the optimal number of nodes for processing S1 images in terms of using the full bandwidth without the risk of saturation can be approximated by the following formula:

$$\begin{aligned} \text{Optimal number of nodes} &= \frac{\text{Processing time}}{\text{I/O time}} = \\ &= \frac{\text{Processing time}}{\text{Read} + \text{Writing time}} = \frac{40 * 60 \text{ sec}}{2 \text{ sec} + 4 \text{ sec}} \cong 400 \text{ nodes} \end{aligned}$$

If less than 400 nodes are used, the bandwidth is not fully exploited while if more than 400 nodes are used the bandwidth will be saturated and computing nodes will automatically be put in a queue by the operative system increasing the chances of a system crash. The same procedure has been employed to ASAR GM and ASAR WS datasets for optimal number of processing nodes evaluation. For ASAR GM the average processing times were less than 1 minute. Groups of 8 images were sent to the computing node for a total of maximum 8 minutes of processing times. Reading and writing operations were performed in less than a second; therefore the optimal number of nodes roughly are:

$$\begin{aligned} \text{Optimal number of nodes} &= \frac{\text{Processing time}}{\text{Read} + \text{Writing time}} = \\ &= \frac{8 * 1 * 60 \text{ sec}}{.5 \text{ sec} + .5 \text{ sec}} \cong 480 \text{ nodes} \end{aligned}$$

For ASAR WS the average processing times were about 5 minutes. As groups of 2 images were sent to a single computing node, 10 minutes of processing times were needed. Reading and writing were performed in less than a second; the optimal number of nodes is therefore:

$$\begin{aligned} \text{Optimal number of nodes} &= \frac{\text{Processing time}}{\text{Read} + \text{Writing time}} = \\ &= \frac{2 * 5 * 60 \text{ sec}}{1 \text{ sec} + 1 \text{ sec}} \cong 300 \text{ nodes} \end{aligned}$$

It is important to note that for ASAR GM and ASAR WS, the strategy to send groups of images has increased the processing time and consequently the number of optimal nodes that are possible to employ.

Table 2. Performance VSC-3 / EODC storage, average experimental values

Performance	VSC-3 BeeGFS distributed storage	EODC storage IBM ESS
I/O rate of 1 CPU writing 1 file of 2.7 GB	1.2 GB/s	1.4 GB/s
Average I/O rate of 5 CPUs each writing 1 file of 2.7 GB	450 MB/s	500 MB/s

All storage parts are accessible from the supercomputer, as well as from the cloud infrastructure, so the final images could be visualized in VMs with a standard GIS software (Figure 3). The technology that has been used for virtualization is OpenStack.

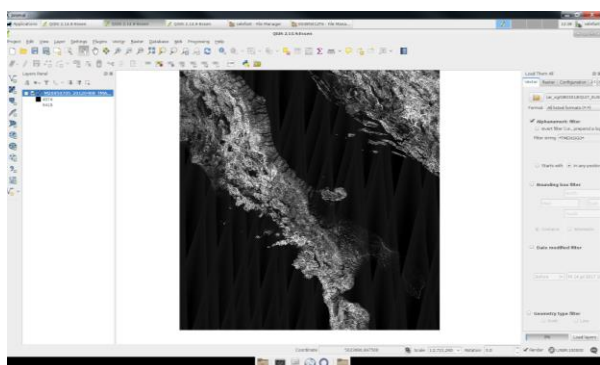


Fig. 3. EODC data visualization using QGIS software

5. CONCLUSIONS

Our experiments demonstrate the feasibility of processing large EO datasets on the EODC high-performance platform. The viability and stability of the platform is evaluated by processing of the whole ENVISAT ASAR GM archive, part of the ASAR WS archive and one large Sentinel-1 data set. The experiments showed that the EODC is a comprehensive infrastructure where users can bring their own algorithm close to EO data, process them by using a supercomputer (VSC-3) and visualize the final products through dedicated VMs, where users can install their own software packages (root privileges granted).

6. ACKNOWLEDGEMENT

This study was supported by the Austrian Research Promotion Agency (FFG) through “Sentinel Big Data Science Cluster” (SBDSC), by ESA and Eumetsat through the projects “EODC business model validation for Exploitation Platforms” and “Satellite Application Facility on Support to Hydrology and Water Management (H-SAF CDOP 3), respectively.

7. REFERENCES

- [1] Hornacek, M., Wagner, W., Sabel, D., Truong, H.L., Snoeij, P., Hahman, T., Diedrich, E. and Doubkova, M., “Potential for High Resolution Systematic Global Surface Soil Moisture Retrieval via Change Detection Using Sentinel-1”, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2012, 5 (4), 1303-1311.
- [2] Pathe, C., Wagner, W., Sabel, D., Doubkova, M. and Basara, J.B., “Using ENVISAT ASAR Global Mode data for surface soil moisture retrieval over Oklahoma, USA”, *IEEE Transactions on Geoscience and Remote Sensing*, 2009, 47 (2), 468-480.
- [3] Vienna Scientific Cluster 3. Available from <<http://vsc.ac.at/systems/vsc-3/>>
- [4] Briese, C., Wagner, W., Boesch, A., Federspiel, C., Aspetsberger, M., Hasenauer, S., Mücke, W., Hoffmann, C., Wotawa, G., “Challenges in the exploitation of big earth observation data”, *Proceeding of the 2014 conference on Big Data from Space*, Frascati (IT)
- [5] BeeGFS. Available from <<http://www.beegfs.com/content/>>
- [6] IBM ESS. Available from <<https://www-03.ibm.com/systems/storage/spectrum/ess/>>
- [7] Naeimi, V., Elefante, S., Cao, S., Wagner, W., Dostalova, A. and Bauer-Marschallinger, B. 2016. Geophysical parameters retrieval from sentinel-1 SAR data: a case study for high performance computing at EODC. In *Proceedings of the 24th High Performance Computing Symposium (HPC '16)*. Society for Computer Simulation International, San Diego, CA, USA, Article 10, 8 pages. DOI: <http://dx.doi.org/10.22360/SpringSim.2016.HPC.026>

SEMANTIC-SENSITIVE HASHING FOR CONTENT-BASED RETRIEVAL IN REMOTE SENSING IMAGES

Thomas Reato, Begüm Demir and Lorenzo Bruzzone

Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
e-mail: reatot@hotmail.it, demir@disi.unitn.it, lorenzo.bruzzone@ing.unitn.it.

ABSTRACT

This paper presents two semantic-sensitive hashing methods that allow fast and accurate retrieval of images in massive remote sensing image archives. The proposed methods aim at mapping high-dimensional image descriptors into multi-hash codes, each of which represents a primitive (i.e., land cover class) present in the images. The first method is unsupervised (i.e., it defines the hash codes by using only the unlabeled images), while the second one is supervised (i.e., it maps the high-dimensional image descriptors to binary codes by exploiting annotated images). Both methods rely on a three-steps algorithm. In the first step, each image is characterized by descriptors of primitives. To this end, images are initially segmented into regions and then descriptors of regions are defined. Then, to define the descriptors of the primitives, regions are associated with primitives present in the images. In the second step, the descriptors of primitives are transformed into multi-hash codes to represent each image. This is achieved by adapting the kernel-based unsupervised and supervised locality sensitive hashing methods to be efficiently used in the proposed unsupervised and supervised multi-code hashing methods, respectively. In the final step, the images in the archive that are similar to a query image are retrieved based on a multi hash code matching scheme. Experiments carried out on an archive of aerial images show that the presented hashing methods significantly improve the retrieval accuracy of the standard hashing methods while keeping the same retrieval time.

Index Terms— remote sensing, content based image retrieval, semantic-sensitive hashing.

1. INTRODUCTION

The continuous advances in satellite technology have led to a significant growth of remote sensing (RS) image archives. Hence, one of the most important research topics is the development of fast and accurate content based image retrieval (CBIR) methods for RS archives. CBIR aims to retrieve relevant images to a given query image from large RS image archives. The simplest approach to CBIR is to exploit the k -nearest neighbor (k -nn) algorithm, which compares the query image with each image in the archive by assessing the similarity among the image descriptors. This algorithm is computationally demanding, when: i) the

number of images in the archive is very high; ii) the considered similarity function is time-demanding to compute; and iii) the dimension of the image descriptor is high [1]. Furthermore, in large-scale CBIR, the storage of the image descriptors in the auxiliary archive (which is a database exploited by performing analyses on the extracted features) is also challenging as RS image contents are often represented in high dimensional descriptors. Thus, the storage of the image descriptors is also a critical bottleneck in RS in addition to the scalability problem [1].

To address these problems, hashing methods have been recently investigated in RS for an accurate and scalable image search and retrieval in large RS data archives [1]. Hashing methods aim at mapping the raw high-dimensional features into binary codes in low-dimensional Hamming space [1]-[3]. Then, similarities of RS images are efficiently measured by simple bit-wise operations, enabling real-time search and accurate retrieval, while reducing the amount of memory required for storing RS image descriptors in the auxiliary archives. To obtain the binary hash codes, hash functions are initially generated and applied to each image descriptor in the archive. Depending on the strategies used to define the hash functions, hashing methods can be unsupervised (hash functions defined by using only the unlabeled images) or supervised (annotated images are used to describe the hash functions).

Kernel-based unsupervised locality sensitive hashing (KULSH) [2] and kernel-based supervised locality sensitive hashing (KSLSH) [3] methods have been recently introduced for RS CBIR problems in [1]. In details, the KULSH defines hash functions in the kernel space by using only unlabeled images, while the KSLSH defines much distinctive hash functions in the kernel space by modeling the semantic similarity by using already annotated images. As any hashing method, these methods characterize each image by a single hash code, which is attained by applying hash functions to global image representations. In other words, they do not model possible primitives present in the images while defining hash functions. However, RS images mostly consist of several regions that can be related to different land-cover classes, representing complex semantic content. Accordingly, characterizing a RS image with a single descriptor, thus with a single hash code, may lead to inaccurate retrieval results, particularly when high-level semantic content is present in the query images. To overcome the limitations of the single hash codes in

complex CBIR tasks, in this paper we introduce semantic-sensitive supervised and unsupervised multi-code hashing methods for large-scale RS retrieval problems.

2. SEMANTIC-SENSITIVE HASHING METHODS

Let $\Upsilon = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_P\}$ be an archive made up of a very large number P of RS images, where \mathbf{X}_i is the i -th image. The proposed methods represent each image with multi-hash codes, each of which corresponds to a primitive in the image. Both methods are defined based on a three-steps algorithm: 1) characterization of each image with descriptors of primitives; 2) transformation of descriptors into semantic-sensitive multi-hash codes; and 3) assessment of the similarities between the multi-hash codes of the query image \mathbf{X}_q ($\mathbf{X}_q \in \Upsilon$ or $\mathbf{X}_q \notin \Upsilon$) and those of each image in the archive Υ to retrieve images similar to the query image. We assume that a set $\mathbf{T} = \{\mathbf{g}_j^T, l_j^T\}_{j=1}^N$ of N annotated training regions with the associated region labels is available for the proposed supervised method. \mathbf{g}_j^T denotes the descriptor of the j -th training region, while $l_j^T \in \mathbf{L}$ indicates its corresponding label and \mathbf{L} is the set of primitive class labels present in the archive. The proposed methods differ from each other on: 1) the considered approach to model the primitives present in the images; and also 2) the considered hashing method for multi-code hashing problems. Each step of the proposed methods is explained in detail in the following by explaining also the main differences between them.

Step 1: Characterization of Images by Descriptors of the Primitives

The first step is achieved by: i) segmenting images $\mathbf{X}_i, i=1, 2, \dots, P$ in the archive into a set of regions $\{r_1^{X_i}, r_2^{X_i}, \dots, r_{n_i}^{X_i}\}$ (where $r_p^{X_i}$ is the p -th region of \mathbf{X}_i and n_i is the total number of regions in \mathbf{X}_i); and ii) computing a region descriptor $\mathbf{g}_p^{X_i}$ (i.e., feature vector that models the region) for each region $r_p^{X_i}, p=1, 2, \dots, n_i$; and iii) associating region descriptors $\mathbf{g}_p^{X_i}$ with primitives present in the images. Primitives are defined based on a clustering based strategy for the unsupervised hashing method, while they are described by exploiting a set of annotated training regions for the supervised hashing method.

In details, associated to the unsupervised method, a set of regions is randomly selected and then clustered into K clusters $\{C_1, C_2, \dots, C_K\}$, where C_k is the k -th cluster. Clustering is achieved by using Gaussian mixture models. Parameters of the mixture models with K components are

estimated by the Expectation Maximization algorithm. Then, the obtained clusters are treated as representative primitives. A correspondence between the regions of an image and the primitives is built based on the probability $P(C_k | r_p^{X_i})$ of each primitive cluster to be present at each region. The posterior probabilities are estimated from the parameters of the mixture models. We consider that all the regions belonging to a specific primitive with a probability higher than a given threshold T are highly representative of that primitive (i.e., cluster). Thus, the average of the descriptors of these regions is used to model the k -th primitive as follows:

$$\mathbf{f}^{X_i, C_k} = \frac{1}{nr} \sum_{\forall P(C_k | r_p^{X_i}) \geq T} \mathbf{g}_p^{X_i} \quad (1)$$

where \mathbf{f}^{X_i, C_k} denotes the descriptor of the k -th primitive in \mathbf{X}_i and nr is the number of regions for which the posterior probabilities are greater than T . If there is no region with posterior probability greater than or equal to the given threshold T (i.e., $P(C_k | r_p^{X_i}) < T, \forall r_p^{X_i}, p=1, 2, \dots, n_i$), we assume that there are not representative regions of the k -th primitive in \mathbf{X}_i . Thus, the related descriptor will be defined as $\mathbf{f}^{X_i, C_k} = \mathbf{z}$, where \mathbf{z} is a vector of all zero entries.

Associated to the supervised method, k -th primitive is represented by a descriptor $\mathbf{f}^{S_k}, k=1, 2, \dots, |\mathbf{L}|$ that is obtained as the average of training region descriptors that belong to the k -th class S_k . Then, unlike the unsupervised method, to build a correspondence between the regions and the primitive descriptors, posterior probabilities $P(S_k | r_p^{X_i})$ of each region $r_p^{X_i}, p=1, 2, \dots, n_i$ within \mathbf{X}_i to belong to the k -th class $S_k, k=1, 2, \dots, |\mathbf{L}|$ are computed as follows:

$$P(S_k | r_p^{X_i}) = \frac{\exp(-\gamma \cdot \|\mathbf{g}_p^{X_i} - \mathbf{f}^{S_k}\|^2)}{\sum_{q=1}^{|\mathbf{L}|} \exp(-\gamma \cdot \|\mathbf{g}_p^{X_i} - \mathbf{f}^{S_q}\|^2)}, \gamma > 0 \quad (2)$$

where γ is the user-defined parameter. Then, to compute the descriptors of the primitives, a similar approach used in the unsupervised method is considered. In other words, (1) is exploited by replacing C_k with S_k .

Step 2: Transformation of the Descriptors of the Primitives into Multi-Hash Codes

In the second step, standard hashing is applied to the descriptors of each primitive independently from each other by properly adapting it to multi hash codes generation

problems. To this end, we consider the kernel-based unsupervised and supervised locality sensitive hashing methods [2], [3] in the framework of proposed unsupervised and supervised hashing methods, respectively. The kernel-based unsupervised locality sensitive hashing (KULSH) method defines the r -th hash function h_r^k of the k -th primitive as follows [2]:

$$h_r^k(\mathbf{f}^{X_i, C_k}) = \text{sign} \left(\sum_{j=1}^P \omega_r^{C_k}(j) K(\mathbf{f}^{X_j, C_k}, \mathbf{f}^{X_i, C_k}) \right) \quad (3)$$

where $\omega_r^{C_k}$ is the coefficient vector and $K(\cdot, \cdot)$ is a kernel function. The same process is applied for a total of b hash functions $[h_1^k, h_2^k, \dots, h_b^k]$, resulting in a b -bits hash code $H_{C_k}^{X_i} = [h_1^k(\mathbf{f}^{X_i, C_k}), h_2^k(\mathbf{f}^{X_i, C_k}), \dots, h_b^k(\mathbf{f}^{X_i, C_k})]$ that characterizes the k -th primitive in \mathbf{X}_i . Then, this is repeated for each primitive and the set of hash codes $\{H_{C_k}^{X_i}\}_{k=1}^{|L|}$ is obtained for each image. The set $\{H_{C_k}^{X_q}\}_{k=1}^{|L|}$ of hash codes of a query image \mathbf{X}_q is also computed by applying the same b hash functions of each primitive.

In our proposed supervised method, the kernel-based supervised locality sensitive hashing (KSLSH) method is used by considering (3) when \mathbf{f}^{X_i, S_k} is replaced with \mathbf{f}^{X_i, C_k} . It is worth noting that the KULSH and KSLSH methods differ from each other on how they estimate the weight vector ω , [2], [3].

It is worth noting that any hashing method that provides single-hash code can be considered to provide semantic-sensitive hash codes after properly applying the first step of the proposed methods.

Step 3: Retrieval of RS Images

To retrieve the images based on the similarities between the multi-hash codes, we exploit a multi hash code matching scheme for both methods. According to this scheme, the similarity between \mathbf{X}_q and $\mathbf{X}_i, i=1, 2, \dots, P$ is estimated as follows:

$$d^{X_q, X_i} = \sum_{k=1}^{|L|} H_{C_k}^{X_q} \otimes H_{C_k}^{X_i} \quad \text{if } \mathbf{f}^{X_q, C_k} \neq \mathbf{z} \quad (4)$$

where d^{X_q, X_i} is the total Hamming distance between the hash codes of \mathbf{X}_q and \mathbf{X}_i , and \otimes represents Boolean exclusive-or operation. Then, the images with the lowest distance d^{X_q, X_i} are finally retrieved.

3. EXPERIMENTAL RESULTS

Experiments were performed on a benchmark archive that consists of 2100 images selected from aerial orthoimagery with a spatial resolution of 30 cm [4]. In the original archive, the images are grouped into 21 different categories as: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks and tennis court. Further details of the archive can be found in [4]. To assess the performance of the proposed methods, we re-annotated the images based on visual inspection and assigned primitive class labels to each of them. The considered primitive class labels are: airplane; bare-soil, buildings, cars, chaparral, court; dock, field, grass, mobile-home, pavement, sand, sea, ship, tanks, trees, water. The total number of labels related to each image is between 1 and 7. The proposed class sensitive supervised hashing method requires a training set of annotated regions. In the experiments, we have used a training set of 9237 regions and their labels, which are representative of the above-mentioned primitive classes. To define the regions of each image, the parametric kernel graph cut segmentation algorithm has been used. After segmenting the images, each region is described by: 1) shape features (which are Fourier descriptors and contour-based shape descriptors); 2) texture features (which are entropy and spectral histograms); and 3) intensity features (which are mean and standard deviation of the samples within each region).

In the experiments, we compared the effectiveness of the semantic-sensitive hashing methods between each other and also with the standard hashing methods (which do not evaluate the primitives present in the images during hashing). Results of each method are analyzed in terms of: i) average recall; ii) storage complexity; and iii) average computational time obtained in 2100 trials performed with 2100 selected query images from the archive.

Table 1 shows the average recall, average computational time and storage complexity required for the proposed methods. From the Table, one can observe that the proposed supervised hashing method provides almost 4% higher recall when compared to the proposed unsupervised hashing method, considering the same storage complexity and taking a slightly higher retrieval time. These results were obtained by considering the total hash bit number as 32 for each primitive. Fig. 1 shows an example of images retrieved by using the proposed unsupervised and supervised hashing methods. The retrieval order of each image is given above the related image, while the primitive class labels associated with the image are given below the related image. By visually analyzing the results, one can observe that the supervised method retrieves semantically more similar images from the archive. The same relative behavior is also

observed in the results obtained by varying the query image (not reported for space constraints).

Table 2 shows the results obtained by proposed and standard unsupervised methods when hash bit number associated to each primitive is 32. From the table, one can see that the proposed unsupervised hashing method provides 6% higher recall when compared to the standard hashing taking the same retrieval time. From our analysis, we have also seen that increasing the number of hash bits leads to higher recall. This is achieved at the cost of increasing the retrieval time and the amount of memory required for storing the hash codes. The similar behavior is also observed in the results obtained by comparing the standard and proposed supervised hashing methods. For space constraints, we could not report these results. All these results show that the multi-hash codes obtained by the proposed methods: 1) efficiently describe the complex content of RS images; 2) allow to achieve scalable image search and retrieval; and 3) overcome the limitations of the single hash codes. Thus, the proposed methods are much more suitable to be used on real RS image retrieval scenarios with respect to the standard hashing methods.

Table 1: Average recall, retrieval time and storage complexity of the proposed unsupervised and supervised hashing methods.

Proposed Methods	Recall	Time (in seconds)	Storage Complexity
Unsupervised	65.29%	41×10^{-4}	0.068 KB
Supervised	69.25%	63×10^{-4}	0.068 KB

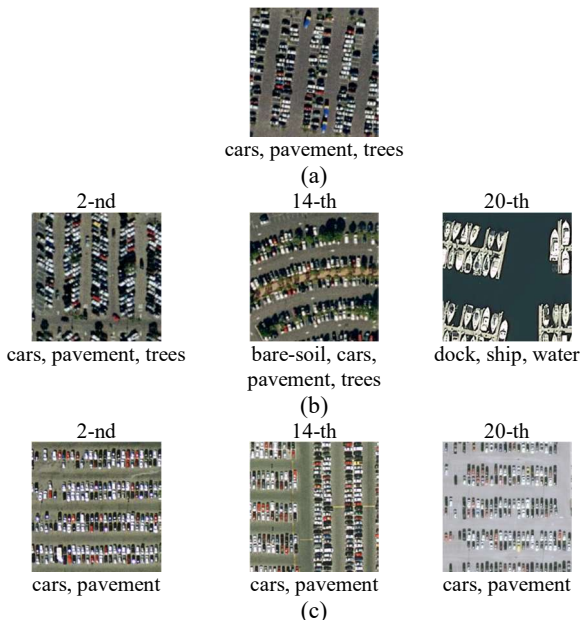


Fig. 1: (a) Query image, (b) images retrieved by the proposed unsupervised hashing method, and (c) images retrieved by the proposed supervised hashing method.

Table 2: Average recall, retrieval time and storage complexity of the standard and the proposed unsupervised hashing methods.

Method	Recall	Time	Storage Complexity
Standard [2]	58.74 %	41×10^{-4}	0.033 KB
Proposed Unsupervised	65.29 %	41×10^{-4}	0.068 KB

4. CONCLUSION

In this paper, we have introduced two semantic-sensitive hashing methods in the framework of content based remote sensing image retrieval. The presented methods allow to achieve scalable image search and retrieval, thanks to mapping high-dimensional image features to multi-hash codes. Moreover, the complex content of RS images is efficiently described by the multi-hash codes that allow improved retrieval performance. Experimental results obtained an archive of aerial images show that the proposed methods are very promising in terms of retrieval time and accuracy, allowing real-time image retrieval in very large RS image archives.

It is worth noting that the proposed methods are associated to different costs for the generation of multi-hash codes. The first method does not need any annotated training regions (i.e., it is fully unsupervised). On the contrary, the second method requires annotated training regions for the generation of multi-hash codes (i.e., it is supervised). However, the hash codes obtained by the proposed supervised method are generally more distinctive than those of the unsupervised method, resulting in higher retrieval accuracy. Note that the choice of the hashing methods may depend on the both availability of the training samples and user's priority on the retrieval accuracy.

As a future development of this work, we plan to apply the proposed hashing methods to the retrieval of long time series of remote sensing images.

5. REFERENCES

- [1] B. Demir, L. Bruzzone, "Hashing based scalable remote sensing image search and retrieval in large archives", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no.2, pp. 892-904, 2016.
- [2] B. Kulis and K. Grauman, "Kernelized locality-sensitive hashing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1092 - 1104, 2012.
- [3] W. Liu, J. Wang, R. Ji, Y. G. Jiang, and S-F. Chang, "Supervised hashing with kernels", *Conference on Computer Vision and Pattern Recognition*, pp. 2074-2081, Rhode Island, USA, 2012.
- [4] Y. Yang, and S. Newsam, "Geographic image retrieval using local invariant features", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818-832, 2013.

SYSTEMATIC ESA EO LEVEL 2 PRODUCT GENERATION AS PRE-CONDITION TO SEMANTIC CONTENT-BASED IMAGE RETRIEVAL AND INFORMATION/KNOWLEDGE DISCOVERY IN EO IMAGE DATABASES

Andrea Baraldi^{1,2}, Dirk Tiede¹, Martin Sudmanns¹, and Stefan Lang¹

¹Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

²Department of Agricultural Sciences, University of Naples Federico II, Portici (NA), Italy

ABSTRACT

In the computer vision (CV) domain, two open problems traditionally coped with independently by the remote sensing (RS) community are: (i) research and development (R&D) of a Global Earth Observation System of Systems (GEOSS), suitable for systematic EO image understanding (EO-IU), and (ii) semantic content-based image retrieval (SCBIR). Our original working hypothesis postulates that: (1) R&D of a CV system in operating mode is necessary not sufficient pre-condition to SCBIR. (2) To become better posed for numerical solution, an inherently ill-posed CV system is constrained to comply with human visual perception. (3) Necessary not sufficient pre-condition to GEOSS is systematic multi-source single-date ESA EO Level 2 product generation. Never accomplished at the ground segment to date, an ESA EO Level 2 product is defined as a single-date multi-spectral image radiometrically corrected for atmospheric, topographic and adjacency effects, stacked with its scene classification map. As proof-of-concept of a GEOSS capable of systematic ESA EO Level 2 product generation necessary to EO-SCBIR, an integrated closed-loop hybrid (combined deductive and inductive) EO-IU for Semantic Querying (EO-IU4SQ) system is proposed as a viable alternative to feedforward inductive learning-from-data CV systems constrained by heuristics, currently dominating the RS and CV literature.

Index Terms—Bayesian inference in vision, cognitive science, European Space Agency (ESA) Earth observation Level 2 product, deductive/inductive/hybrid inference, semantic content-based image retrieval, world model.

1. INTRODUCTION

The visionary goal of a Global Earth Observation System of Systems (GEOSS) implementation plan for years 2005-2015 proposed by the intergovernmental Group on Earth Observations (GEO) was systematic transformation of multi-source Earth Observation (EO) “big data” into timely, comprehensive and operational EO value-adding information products and services [1]. Hereafter, the popular GEOSS acronym is considered synonym (actually, *superset-of*) of EO image understanding (EO-IU) in operating mode, i.e., $GEOSS \approx EO-IU$. To be considered in operating mode, an EO information processing system must score “high” in each

index of a minimally dependent and maximally informative (mDMI) set of outcome and process quantitative quality indicators (OP-Q²Is), in agreement with the GEO’s Quality Assurance Framework for Earth Observation (QA4EO) calibration/validation (*Cal/Val*) requirements specification [1]. Proposed EO OP-Q²Is are [2], [3]: (i) degree of automation, (ii) accuracy, provided with a degree of uncertainty in measurement, (iii) efficiency, (iv) robustness to changes in input parameters, (v) robustness to changes in input data, (vi) scalability to changes in input sensor specifications and user requirements, (vii) timeliness from data acquisition to product generation and (viii) costs in manpower and computer power. Unfortunately, EO-IU systems presented in the remote sensing (RS) literature are typically assessed and compared based on the sole mapping accuracy without degree of uncertainty.

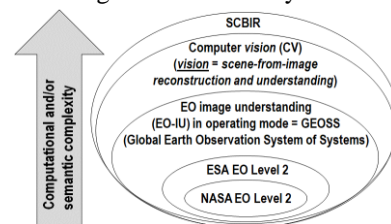


Fig. 1. Our original working hypothesis postulates that: human vision (works as lower bound of) \rightarrow ESA EO Level 2 product generation, whose special case is NASA EO Level 2 product generation \rightarrow GEOSS = EO-IU \rightarrow CV \rightarrow (EO-)SCBIR, where symbol ‘ \rightarrow ’ denotes relationship part-of pointing from the supplier to the client according to the UML notation.

Based on several true-facts, no GEOSS can be considered accomplished by the RS community to date. One observation is that no systematic European Space Agency (ESA) EO Level 2 product generation at the ground segment has ever been fulfilled. For example, the Sentinel-2 data Correction (SEN2COR) Prototype Processor recently developed and distributed by ESA must be run on user side [4]. By definition an ESA EO Level 2 product comprises [4]: (i) a single-date multi-spectral (MS) image corrected for atmospheric, adjacency and topographic effects, stacked with (ii) its data-derived general-purpose, user- and application-independent scene classification map (SCM). Since the definition of NASA EO Level 2 product is “data-derived geophysical variable at the same resolution and location as Level 1 source data” [2], dependency NASA EO Level 2 \rightarrow ESA EO Level 2 holds true, where symbol ‘ \rightarrow ’ denotes

relationship *part-of*, pointing from the supplier to the client according to the Unified object Modeling Language (UML) notation (Fig. 1).

Another unquestionable true-fact proving the lack of GEOSS is that no EO semantic content-based image retrieval (SCBIR) system, capable of semantics-enabled knowledge/information discovery in EO image databases [5], has ever been developed in operating mode to process semantic queries such as “retrieve all EO images not necessarily cloud-free acquired by sensor X where wetland is visible and located adjacent to a highway near the eastern cost of country Y” [5]. It means that prototypical EO content-based image retrieval (CBIR) systems, capable of sub-symbolic (non-semantic) image querying by metadata text information including image-wide summary statistics (e.g., per-image cloud cover quality index computed off-line), and/or by either image, image-object or multi-object examples, support no SCBIR capability because their EO-IU subsystem, if any, scores low in operating mode.

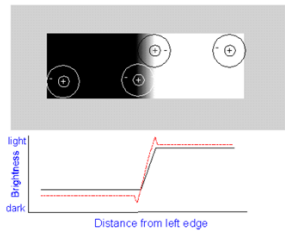


Fig. 2. Mach bands illusion: where a luminance (intensity) ramp meets a plateau, there are spikes of brightness (perceived luminance), where humans detect ramp edges, although there is no discontinuity in the luminance profile, independent of its slope. This simple, but not trivial perceptual effect is at odd with a large portion of existing semi-automatic image segmentation/image-contour detection algorithms based on heuristic thresholding of local variance, contrast or gradient statistics [6].

Vision, encompassing both biological and computer vision (CV), is synonym of scene-from-image reconstruction and understanding. This is an inherently ill-posed cognitive task affected by, first, data dimensionality reduction from the 3D scene-domain to the (2D) image-domain and, second, by a semantic information gap from ever-varying sensations (sensory data) in the image-domain to stable percepts (concepts) in the mental model of the real-world (world-model, world ontology). Both GEOSS \approx EO-IU in operating mode and SCBIR pertain to the cognitive domain of CV. Traditionally the RS community has coped with the ambitious goals of GEOSS and SCBIR as independent tasks. On the contrary, our original working hypothesis postulates: human vision \rightarrow ESA EO Level 2 product generation \rightarrow GEOSS \approx EO-IU \rightarrow CV \rightarrow (EO)-SCBIR, with symbol ‘ \rightarrow ’ denoting dependence relationship *part-of* pointing from the client to the supplier (Fig. 1). Its meaning is threefold, as reported in the *Abstract*. For example, an inherently ill-posed CV system is constrained to comply with human visual perception to become better-posed for numerical solution.

According to Pessoa, “if we require that a CV system should be able to predict perceptual effects, such as the well-known Mach bands illusion where bright and dark bands are seen at ramp edges (Fig. 2), then the number of published vision models becomes surprisingly small” [2]. In fact, well-known image-contour detection and image segmentation algorithms, adopted by popular open-source and commercial RS image processing software toolboxes and based on thresholding local variance, contrast or gradient statistics [6], are inconsistent with the Mach bands perceptual illusion when coping with ramp edges.

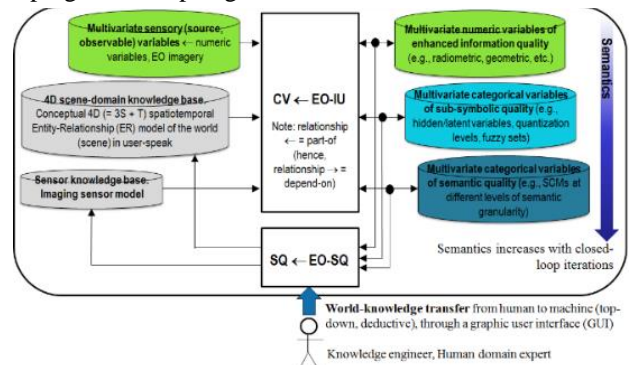


Fig. 3. Closed-loop EO-IU4SQ system architecture, suitable for incremental semantic learning. It comprises a primary (dominant) hybrid feedback EO-IU subsystem in closed-loop with a secondary (dominated) hybrid feedback EO-SQ subsystem.

As proof-of-concept of a GEOSS capable of systematic ESA EO Level 2 product generation necessary to EO-SCBIR, we propose the R&D of an integrated hybrid (combined deductive and inductive) closed-loop (feedback) EO Image Understanding for Semantic Querying (EO-IU4SQ) system, consisting of: (A) a primary (dominant, necessary not sufficient) hybrid feedback EO-IU subsystem, capable of automated near real-time multi-source single-date ESA EO Level 2 product generation as initial condition, equivalent to a Bayesian prior in a Bayesian inference approach to CV [2]. (B) A secondary (dependent) hybrid feedback EO Semantic Querying (EO-SQ) subsystem, capable of incremental semantic learning/querying starting from ESA EO Level 2 product instantiations as initial condition (Fig. 3). The proposed hybrid closed-loop EO-IU4SQ system is alternative to feedforward inductive learning-from-data CV systems constrained by heuristics, such as deep convolutional neural networks (DCNNs), which are currently dominating the RS and CV literature.

Awarded with the *Copernicus Masters 2015 T-Systems Big Data Challenge*, a prototypical implementation of a hybrid closed-loop EO-IU4SQ system has been developing since 2015, as reported below.

2. MATERIALS AND METHODS

In the EO-IU4SQ system, the primary EO-IU subsystem accomplishes systematic automated (without human-

machine interaction) near real-time (in linear complexity with image size) multi-source single-date ESA EO Level 2 product generation. To show its degrees of novelty, it is compared against the existing Sentinel-2 image-specific ESA SEN2COR prototype processor to be run on user-side [4]. Levels of CV system comparison are [2]: (i) knowledge/information representation, (ii) system architecture (design), (iii) algorithm, and (iv) implementation, which are discussed hereafter.

	Pseudocolor
A11	1. Cultivated and Managed Terrestrial (non-aquatic) Non-vegetated Areas
A12	2. Natural and Semi-Natural Terrestrial Vegetation
A23	3. Cultivated Aquatic or Regularly Flooded Vegetated Areas
A24	4. Natural and Semi-Natural Aquatic or Regularly Flooded Vegetation
B35	5. Artificial Surfaces and Associated Areas
B36	6. Bare Areas
B47	7. Artificial Waterbodies, Snow and Ice
B48	8. Natural Waterbodies, Snow and Ice.
	9. Quality layer: Cloud
	10. Quality layer: Cloud-shadow
	11. Others

Fig. 4. “Ideal” standard general-purpose, user- and application-independent 3-level 8-class FAO LCCS Dichotomous Phase (DP) taxonomy [7], “augmented” with quality layers cloud and cloud-shadow + class Others (Unknowns). The 3-level hierarchical FAO LCCS-DP mapping criteria are: (i) vegetation versus non-vegetation, (ii) terrestrial versus aquatic, (iii) managed versus natural or semi-natural, generating the eight LC classes identified as A11 to B48 [7].

(i) Knowledge/information representation. The standard fully nested 3-level 8-class FAO Land Cover Classification System (LCCS) Dichotomous Phase (DP) taxonomy [7], augmented with quality layers cloud and cloud-shadow (Fig. 4), is proposed as sensor-, user- and application-independent ESA EO Level 2 SCM legend. Its semantics is far superior than that of the non-standard SEN2COR taxonomy (Fig. 5).

Label	Classification
0	NO_DATA
1	SATURATED_OR_DEFECTIVE
2	BARK_AREA_PIXELS
3	CLOUD_SHADOWS
4	VEGETATION
5	BARE_SOILS
6	WATER
7	CLOUD_LOW_PROBABILITY
8	CLOUD_MEDIUM_PROBABILITY
9	CLOUD_HIGH_PROBABILITY
10	THIN_CIRRUS
11	SNOW

Fig. 5. General-purpose, user- and application-independent ESA Level 2 SCM’s legend adopted by the Sentinel-2 image-specific SEN2COR prototype processor, to be run on user side [4].

(ii) System architecture. To transform MS dimensionless digital numbers (DNs) into a sequence of top-of-atmosphere radiance (TOARD), top-of-atmosphere reflectance (TOARF) and surface reflectance (SURF) values, the proposed sensor-independent ESA EO Level 2 product generator adopts a hierarchical workflow of MS image classification stages alternated with driven-by-prior-knowledge (stratified, masked) MS image radiometric enhancement steps (Fig. 6). Inherently ill-posed (chicken-and-egg dilemma) topographic correction and atmospheric correction problems, require input data stratification principles to become better conditioned for numerical solution [3]. In SEN2COR, only one SCM map is generated from TOARF values, based on a

per-pixel (spatial context-insensitive) prior spectral knowledge-based decision tree, at the beginning of the atmospheric, topographic and adjacency correction sequence.

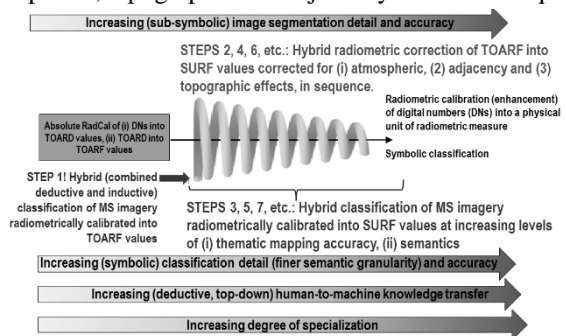


Fig. 6. Proposed ESA EO Level 2 product generator as a hierarchical alternating sequence of MS image classification and driven-by-knowledge (stratified) MS image radiometric enhancement steps, to make inherently ill-posed topographic correction and atmospheric correction problems better posed for numerical solution [3].

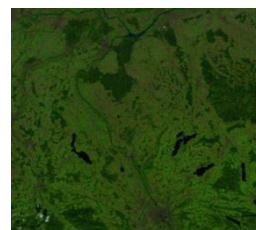


Fig. 7(a). 12-band Sentinel-2 image of Austria, radiometrically calibrated into TOARF values. Depicted in false colors (R = MIR1, G = NIR, B = Visible Blue), 10 m resolution. No histogram stretching for visualization purposes

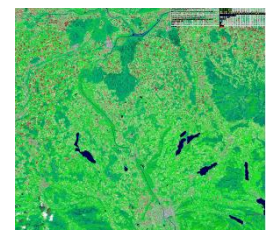


Fig. 7(b). SIAM color map at fine color granularity, consisting of 96 spectral categories depicted in pseudo colors. Color map legend:

(iii) Algorithms. (a) Deductive (prior knowledge-based) color naming, equivalent to static (non-adaptive to data) hyperpolyhedralization of a MS reflectance space [2]. For color naming, the proposed sensor-independent ESA EO Level 2 product generator adopts the Satellite Image Automatic Mapper (SIAM) [2], [3]. SIAM is a lightweight computer program for automated near real-time multi-sensor MS reflectance space hyperpolyhedralization into a static dictionary of MS color names by means of a physical model-based spectral decision tree, followed by superpixel/texel (connected sets of pixels featuring the same color name) detection and vector quantization (VQ) quality assessment (Fig. 7). Unfortunately, MS reflectance space hyperpolyhedra for color naming are difficult to think of and impossible to visualize when the MS data space dimensionality is superior to three. Spectral rule-based decision trees implemented by SIAM and SEN2COR differ at the implementation level. SIAM adopts a fuzzy convergence-of-evidence approach to model families of spectral signatures, where fuzzy evidence from multi-variate spectral shape and multi-variate spectral intensity information components are combined [2], [3]. On

the contrary, SEN2COR adopts crisp thresholding of univariate (scalar) spectral indexes. Any spectral index is equivalent to an angular coefficient of a tangent to the spectral signature in one point. Infinite functions can feature the same tangent in one point. Hence, spectral index thresholding is never robust to changes in input data. (b) To make the inherently ill-posed topographic correction problem better posed for numerical solution, a SIAM-driven and digital surface model (DSM)-based stratification of the topographic corrector is implemented as proposed in [3]. This physical model-based input data stratification strategy is more effective (informative), efficient and accurate than the heuristic data stratification adopted in SEN2COR. (c) Stratified atmospheric correction, where SIAM replaces the oversimplistic spectral rule-based decision tree adopted by NASA LEDAPS [2]. (d) Cloud/cloud-shadow detection. A hybrid (combined deductive and inductive) spatial context-sensitive cloud/cloud-shadow detector proposed in [2] is implemented as a viable alternative to the prior knowledge-based (deductive) per-pixel (spatial context-insensitive) cloud detector implemented in SEN2COR. (e) Land cover (LC) class classifier. Alternative to the Sentinel-2 pixel-based static decision tree implemented in SEN2COR whose taxonomy is shown in Fig. 5, the proposed multi-sensor LC classifier, whose taxonomy is shown in Fig. 4, adopts a hybrid spatial-context-sensitive and spatial topology-preserving convergence-of-evidence approach, where visual information sources are: (i) color names identified by SIAM, (ii) local shape of planar objects, (iii) texture (perceptual spatial grouping of texture elements, texels), and (iv) inter-object spatial relationships, either topological (e.g., adjacency, inclusion, etc.) or non-topological [2]. Examples of multi-source ESA EO Level 2 SCM instances automatically generated in linear complexity with image size are shown in Fig. 8 to Fig. 10. Examples of semantic queries supported by the EO-SQ subsystem's graphic user interface can be found in [2].

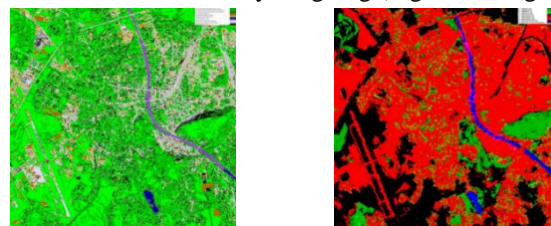


Left, Fig. 8(a). 4-band (B, G, R, NIR) ALOS AVNIR-2 image of Campania, Italy, radiometrically calibrated into TOARF values. Depicted in false colors (R = Visible Red, G = NIR, B = Visible Blue), 10 m resolution. No histogram stretching for visualization purposes. Right, Fig. 8(b). ESA EO Level 2 SCM classification map. Map legend shown in Fig. 4.

3. CONCLUSIONS

Neither GEOSS nor multi-source ESA EO Level 2 product generation at the ground segment nor SCBIR in operating mode have ever been accomplished by the RS community to

date. A hybrid closed-loop EO-IU4SQ system design and implementation are proposed as proof-of-concept of a GEOSS where systematic automated near real-time ESA EO Level 2 product generation is accomplished as pre-condition to EO-SCBIR, according to Bayesian inference in CV. In agreement with the QA4EO *Cal/Val* requirements [1], validation by independent means of generated ESA EO Level 2 SCM instances is currently on-going (Fig. 9 and Fig. 10).



Left, Fig. 9(a). Subset of an ESA EO Level 2 SCM map automatically generated from the Sentinel-2 image shown in Fig. 7(a). Map legend shown in Fig. 4. Right, Fig. 9(b). Reference thematic map. European Environment Agency (EEA), GIO Land (GMES/Copernicus Initial Operations Land), Pan-European components: High Resolution Layers (HRLs), reference year 2012, 20 m spatial resolution, upscaled to 10 m. Map legend shown in Fig. 10.

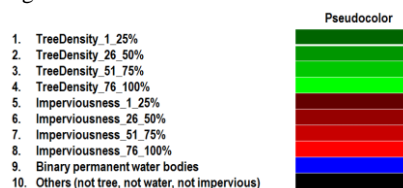


Fig. 10. Reference EEA GIO Land, Pan-European components: HRLs, reference year 2012, 20 m spatial resolution. Revisited map legend: 4 fuzzy sets {1-25, 26-50, 51-75, 76-100} in Tree Density [0%, 100%], 4 fuzzy sets {1-25, 26-50, 51-75, 76-100} in Imperviousness Density [0%, 100%], Binary permanent water bodies.

4. REFERENCES

- [1] Group on Earth Observation (GEO) (2010). Quality Assurance Framework for Earth Observation 4.0. http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf
- [2] Baraldi, A. (2017). Pre-processing, classification and semantic querying of large-scale Earth observation spaceborne/airborne/terrestrial image databases: Process and product innovations, Univ. Naples Federico II, Ph.D. dissertation.
- [3] Baraldi, A., Girona, M., & Simonetti, D. (2010). Operational three-stage stratified topographic correction of spaceborne multi-spectral imagery employing an automatic spectral rule-based decision-tree preliminary classifier. *IEEE Trans. Geosci. Remote Sensing*, 48(1), 112-146.
- [4] DLR and Telespazio VEGA. (2011). "Sentinel-2 MSI – Level 2A Products Algorithm Theoretical Basis Document.", ESA.
- [5] Dhurba, S., and King, R. (2005). Semantics-enabled framework for knowledge discovery from Earth observation data archives. *IEEE Trans. Geosci. Remote Sens.*, 43(11), 2563-2572.
- [6] Baatz, M. & Schäpe, A. (2000). Multiresolution Segmentation: An Optimization Approach for High Quality Multi-Scale Image Segmentation. Verlag: Berlin, Germany, 58, 12–23.
- [7] Di Gregorio, A., & Jansen, L. (2000). Land Cover Classification System (LCCS). FAO: Rome, Italy.

DATA CORRELATION: FUSING LOGFILES WITH PERFORMANCE COUNTERS TO DIAGNOSE PERFORMANCE ISSUES IN GROUND SYSTEMS

Paschalis Veskos¹, Stathis Koukouvinos¹, Fabien Castel²

¹Software Competitiveness International, Athens Greece,

²Atos Integration, Toulouse France

ABSTRACT

Systems produce functional log files during their execution. At the same time, the resources consumed on a server can be monitored during this execution. The main idea behind this study is to correlate the functional logs with resource consumption measurements to assess the performances of a system. This is done using the Syer et al. algorithm [1], which allows diagnosing memory related performance issues from execution logs and performance counters. This was previously implemented as a desktop application for ESA under the project “Leveraging system performance metrics and execution logs to proactively diagnose system of systems performance issues” [2]. We address the reimplementation of this algorithm for a big-data use case, applied to ground systems for the space industry.

Index Terms— Enterprise application performance monitoring, load testing, correlation engine, Apache Spark, Hadoop, Big Data, machine learning, GAIA mission

1. INTRODUCTION

Traditional data mining approaches process log files or application performance counters to extract information and present users with higher-level knowledge about application behavior. The Syer algorithm differentiates itself from these approaches in that it combines these two sources of information, looking for correlations between them. This way it can diagnose performance issues by attributing anomalies recorded in the performance counters to specific log lines present in log files. It is thus categorized as a *correlation engine*.

This usage scenario then assumes that a) both log files and performance counters are available to the application and b) that a performance issue has already been detected by an operator and it is the cause that is being investigated, not its existence (i.e. performs issue *diagnosis*, not detection).

The Syer algorithm consists of 7 phases and 3 processing steps (Figure 1). The first two phases belong to the data preparation step, where log lines are abstracted into execution events and along with the performance counter samples are discretized into time-slice profiles. These describe the log activity and its corresponding impact on the memory consumption. The duration of the timeslices is

determined by a preset sampling interval. The next step is the clustering procedure and consists of three phases. Here the profiles are clustered into groups where similar log activity has occurred. Using a dendrogram, this stage outputs the optimal clustering of profiles. The final step is cluster analysis, where the statistics of the clusters are inspected to find outliers (clusters with high impact on memory consumption). The events belonging to these outlying clusters are identified in the influence analysis step and are output as anomalous events: those considered most likely to be the cause of the memory issue at hand.

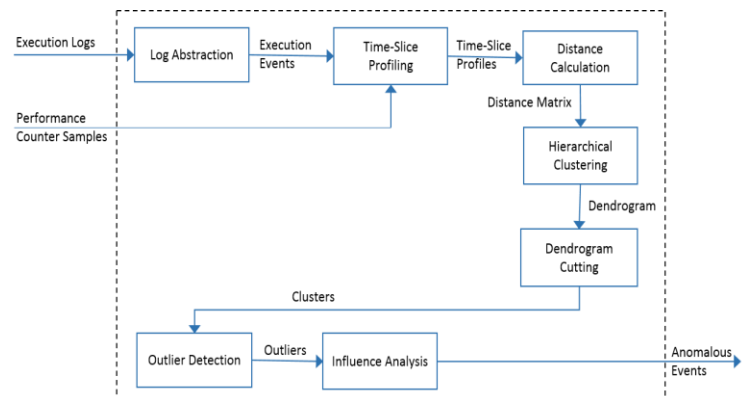


Fig. 1 Syer algorithm overview

2. IMPLEMENTATION

During this study, it became clear that the existing desktop Java implementation was not suitable for processing the very large files produced in 24-hour load tests by modern ground systems. This was for two main reasons: a) very large input datasets would create intermediate data structures during processing whose sizes would exceed the maximums allowed by the Java language and b) certain code paths were too computationally expensive to be useable at this scale.

To solve both problems, the entire algorithm was reimplemented using the distributed data structures and parallel processing patterns of the Apache Spark cluster computing framework [3]. The distributed implementation would allow the algorithm to scale horizontally by adding more processing nodes to clusters, deploying to the cloud if

necessary. Additionally, the intermediate data structures produced during processing can make use of the memory available to several machines in the computing cluster.

Most algorithm phases could be straightforwardly parallelized, greatly benefitting from the framework's scalability. The most computationally expensive phase was hierarchical clustering that, as an iterative process, data-dependent on previous results does not easily parallelize. To overcome this limitation, we made use of a recently published, alternative parallel hierarchical clustering algorithm [4] that uses Apache Spark. This algorithm offers the same time complexity as the serial version, but can scale linearly with multiple processing cores.

We noted however that this approach outputs the result in a different format requiring further iterative processing that can only be performed serially (a minimal spanning tree that needs to be converted to the final dendrogram). Thus the process can only be partially parallelized; this is a known problem for hierarchical clustering and remains a limitation of the Syer algorithm.

The application tries to find the anomalous event, so we provide the ranking of the real event as a measure of algorithm accuracy. Ideally, it should calculate the highest confidence value for the actual anomalous event, so we provide both the ranking and the percentage confidence to indicate the certainty the algorithm has for this calculation. This information can be used by analysts to gauge the importance of each identified event. The application also outputs the precision (the ratio of true positives to total identified anomalous events, including false positives) and recall (the fraction of true positives to total relevant events, including false negatives) as statistical benchmarks for accuracy.

The previous study identified cases where the final algorithm phase (influence analysis) did not work properly, missing certain overly influential events with many occurrences or events with big distance changes during the analysis process. For this reason, we implemented an alternative analysis algorithm that extends the results list with this kind of events. This improvement led to more anomalous events being identified, but sometimes it also helped the identification of normal events (false positives). The purpose of the correlation engine is to point out to the analyst some anomalous events to examine as a starting point for debugging the application. Since the correlation process would otherwise have to be performed manually, the incorporation of a small number of false positives might not be a severe problem. The alternative algorithm can be enabled and disabled at the user's request.

3. ALGORITHM TESTING

In order to generate realistic test datasets for the algorithm, we used process-model generators to simulate enterprise

application behavior. These created both transient (memory spikes) and persistent (memory leaks) performance issues simultaneously generating suitable execution logs. The former simulated a real-world Java application causing configurable memory spikes and the latter a real-world C++ application causing configurable memory leaks. Lightweight instrumentation was used to collect performance counters at the system level, respectively using JMX to get information from the JVM and DTrace to query the operating system kernel.

All log files and performance counters were collected for 24-hour long test periods. Systematically varying process parameters allowed evaluation of their effects on algorithm accuracy and performance.

For a large set of test file parameter combinations, the algorithm correctly identified the anomalous event in all cases when searching for memory spikes with a high confidence value. The leak event was identified most of the times with the highest confidence value. The single time that it was second, the difference in confidence values from the first was small.

To test the parallelization efficiency (horizontal scalability) of using the Spark framework, the prototype was deployed on different size computing clusters (8 & 64 cores) and compared to single machine multicore execution (Spark local mode on a 2-core machine). We ran the algorithm repeatedly with smaller sampling intervals to benchmark the execution time and the results are shown in Figure 2. As expected, the code scaled well with more available CPU resources and can process files far larger than the single threaded desktop version is capable of. For files small enough to be processed by the desktop algorithm version, processing time is dominated by network latency and the parallel version is much slower.

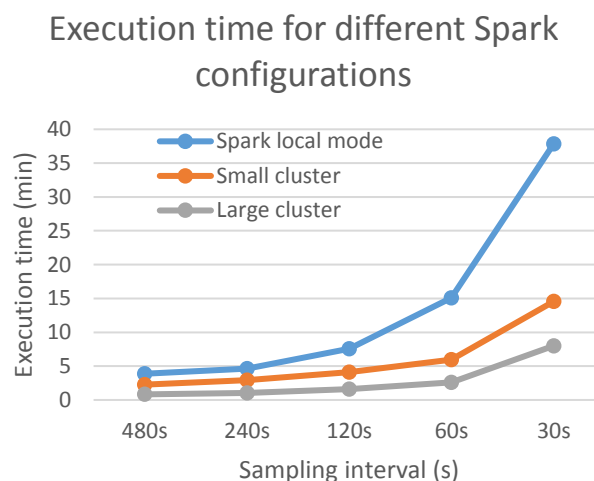


Fig. 2 Execution times for the same file for different Spark configurations & sampling intervals (on 2, 8 & 64 cores).

To test the scaling performance of the application on large input datasets, we generated a number of large test files and note the execution time of the algorithm. To increase file size, we used a large number of concurrent processes generating loglines (log sizes 5-16Gb). We processed the files on the large Spark cluster with a low 30s and a higher 60s sampling interval to increase processing time and stress test the application. For these runs, the application used almost all of the 40Gb RAM per processing node on the large cluster (Spark memory 160Gb).

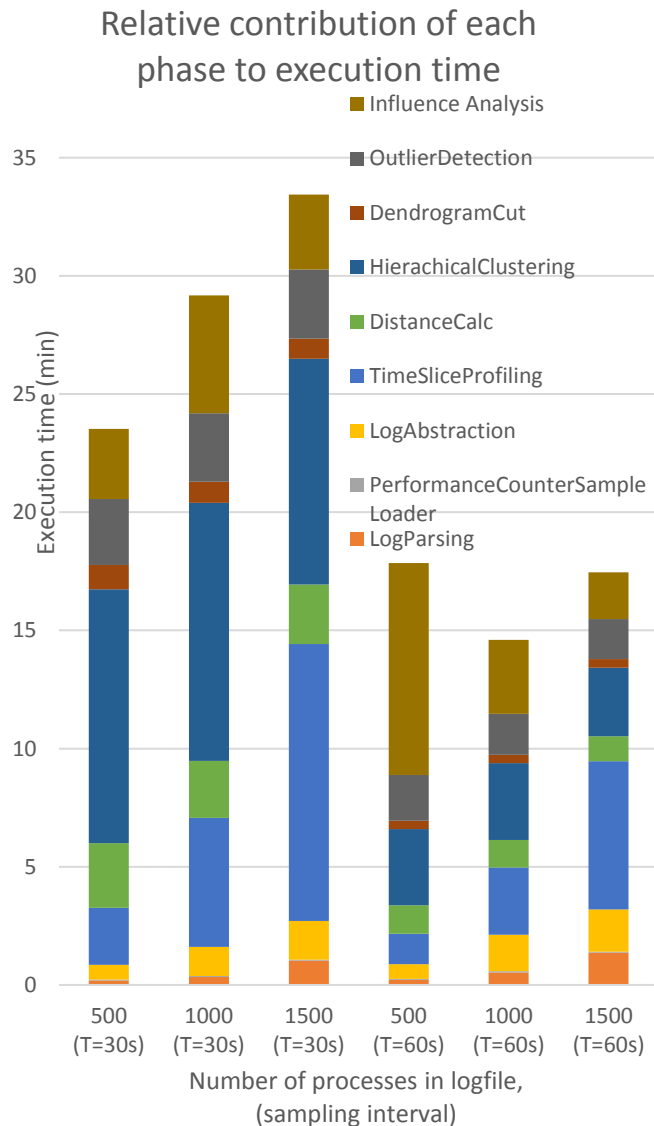


Fig. 3 Processing times breakdown per algorithm phase for large input datasets for two different sampling intervals.

We found that for a given number of timeslice profiles (load test duration) processing time does not vary greatly with increasing log file size. We noted however, that for the 60s sampling interval case the smaller file takes longer to process. To investigate this further, we noted the processing time needed for each algorithm phase. This breakdown is

shown in Figure 3. For the 30s case, processing time was dominated by the hierarchical clustering phase, which remained constant across the 3 tests, since they all contained the same number of samples and hence the number of profiles generated is fixed. Increasing log file size greatly increased the importance of timeslice profiling and most of the processing time increase can be attributed to this phase.

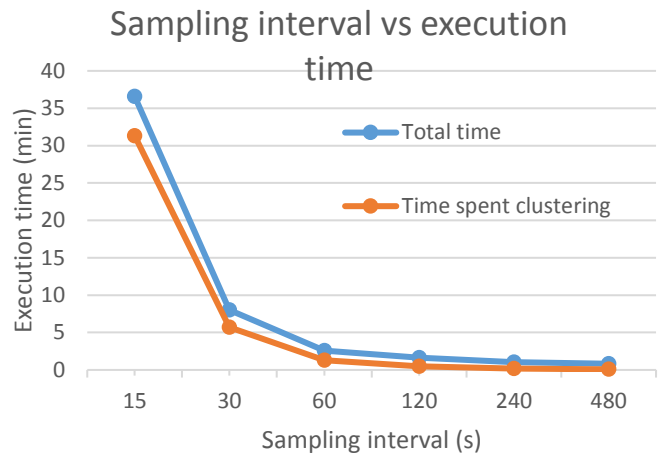


Fig. 4 Effect of decreasing the sampling interval on the algorithm execution time.

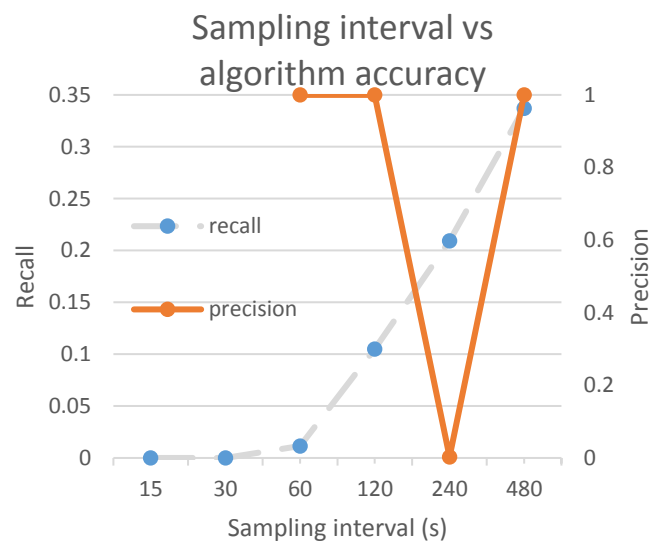


Fig. 5 Changing the sampling interval has an adverse effect on the accuracy of the algorithm (It is not possible to calculate precision for the lowest sampling interval values since no spikes are found)

For the 60s case however, the influence analysis phase made a disproportionately large contribution to the total. This is due to the time complexity of the last two phases depending on the specific input data: The number of outliers (and hence influencers) found in the dataset can vary. This makes execution duration non-deterministic, unlike other phases that only depend on the number of loglines or profiles.

We also found that abnormally reducing the sampling interval of the algorithm had a detrimental effect on execution time and algorithm accuracy (Figures 4 & 5). This made using the algorithm with a 1-second sampling interval, as initially requested, not practical. While this may appear to be counterintuitive, it is consistent with the original Syer approach that used much larger sampling intervals of the order of 60-180s.

We also tested other parameters not present in the single-threaded version of the algorithm: The number of splits controls the number of partitions the profiles are split into during the clustering phase and does not affect algorithm accuracy, but requires tuning to find the best parallelization. Trying values of 2, 5, 10 & 20, the lowest execution time was for 10 splits, as further increasing this value did not help the algorithm parallelize further.

Lastly, it is worth mentioning that the use of the single linkage criterion (required in the distributed hierarchical clustering phase instead of the average linkage in the serial version) did not affect the accuracy of our results.

4. VALIDATION

This study goal is to validate the use of such approach with a real system and not only with theoretical or simulated systems. A scientific processing chain called GAIA, operated by the CNES, was used for the real-life validation of the correlation engine. It is a distributed system able to provide a wide range of monitoring and logging data that could be used to define a realistic analysis test case for the Syer algorithm. GAIA provides various monitoring data (system and application logs, performance measurements...) all gathered in structured and easy to parse files. Dataset were provided by CNES and adapted to be exploitable by the correlation engine. Studying the data, some limitations were highlighted, the most important one being that the performance measurements are recorded only on a small subset of the machines hosting the GAIA system, and not all of the time. Due to these limitations in available data, a multi-host scenario could not be built in the immediate context of the study. Instead a single host basic scenario was run on the subset of exploitable data. The obtained results were promising even if they would need to be validated by a GAIA expert and should be put in perspective with the small amount of analyzed data

5. CONCLUSIONS & FURTHER WORK

The reimplementing of the Syer algorithm using the Apache Spark distributed processing framework led to a prototype that scales well horizontally, making good use of the resources available in large computing clusters. It was possible to parallelize most algorithm phases, greatly accelerating performance. The most expensive operations however, could only be partially parallelized. The serial

execution code paths in them remain as the main execution bottleneck.

Similarly, in terms of space, the prototype is capable of processing much larger input files at greater granularity than the original. However, the requirement of implementing certain phases in a particular manner creates certain blocks to further scaling up the prototype. The main issue is the all-to-all calculation required for the as the distance calculation phase that can get very expensive in terms of memory required to complete it.

This way we were not capable of scaling the implementation to the initial estimation of processing hundreds of terabytes of log files for a real-life test case. These numbers however, were greatly revised downwards during the project and our implementation is more than capable of processing files at the order of magnitude requested (tens of gigabytes on a 64-cpu cluster).

To further improve the performance of the prototype, it could be envisaged that the algorithm is altered to not require these exact processing steps. This would make sense as the Syer algorithm is not optimized for production, but is rather part of a research project whose main advantage is the novel approach of not only processing log files as traditional data mining does, but combining them with performance counters in an automatic manner.

The other issue that could be improved would be to test the prototype's accuracy on more real-world data. This proved to be particularly difficult due to the lack of availability of performance counters. Unlike log files, these are normally not persisted in general use and there is no repository of datasets of performance counters collected during typical load tests. For this testing to be meaningful, there would have to be more involvement from the operators of the systems that are going to be monitored so that their input can be used to improve the prototype's accuracy.

6. REFERENCES

- [1] M.D. Syer, Z.M. Jiang, M. Nagappan, A.E. Hassan, M. Nasser & P.Floria, Leveraging Performance Counters and Execution Logs to Diagnose Memory-Related Performance Issues, *2013 IEEE International Conference on Software Maintenance (ICSM)*, pp. 110-119
- [2] ESA Study "Leveraging System Performance Metrics and Execution Logs to Proactively Diagnose System of Systems"
- [3] M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker, & I. Stoica. "Spark: Cluster Computing with Working Sets." *HotCloud 10* (2010): 10-10.
- [4] C. Jin, Md. Mostofa, A. Patwary, A. Agrawal, W. Hendrix, W. Liao & A. Choudhary, "DiSC: A Distributed Single-linkage Hierarchical Clustering Algorithm using MapReduce", *International Workshop on Data Intensive Computing in the Clouds (DataCloud)*, 11/2013.

ON-BOARD DEBRIS DETECTION BASED ON GPU TECHNOLOGY

Francesco Diprima ^{(1)(3)*}, Fabio Santoni ⁽²⁾, Fabrizio Piergentili ⁽¹⁾,
Vito Fortunato ⁽³⁾, Cristoforo Abbattista ⁽³⁾, Leonardo Amoroso ⁽³⁾

⁽¹⁾DIMA, Sapienza University of Rome, Via Eudossiana 18, Rome, Italy

⁽²⁾DIAEE, Sapienza University of Rome, Via Eudossiana 18, Rome, Italy

⁽³⁾Planetek Italia s.r.l., Via Massaua 12, Bari, Italy

* Corresponding author

ABSTRACT

The presence of an enormous number of space debris (SD) in the Earth orbit it is a serious threat for the active satellites. Agencies and companies that manufacture satellites needed a monitoring system for preventing collisions.

The estimate of the collision risk is calculated based on the SD orbital propagation. To obtain accurate orbit it is necessary use real data. The optical data allows calculating the objects' angular position comparing the stars inside the image with stars catalogues by astrometry procedures.

We present a system based on a nearly continuous scanning of the sky, able to collect a significant number of observations. Core of the system is the on-board processing able to detect in automatic the SD in optical data using image-processing techniques.

Considering the huge amount of collected data, we also examine the advantages of increasing the performance exploiting parallelism by means of Graphics Processing Units (GPU).

Index Terms— Space debris, Automated image analysis, Object detection, GPU

1. INTRODUCTION

With the development of technology and space launches, the space debris (SD) present in orbit has increased vastly, becoming a serious threat to space activity [1]. Computer models based on observations of SD are used to predict future growth of the SD. These models show that in future the SD population in the LEO orbit would reach a critical point where the rate of growth would be greater than the rate at which these would be removed from orbit through natural decay of the Earth's atmosphere [2]. To guarantee safe use of LEO orbits, Active Debris Removal (ADR) systems will have to be implemented.

Anyway, a SD monitoring is a fundamental point to support any future activity of debris removal.

SD monitoring activities are currently carried out by means of a ground network of optical and radar space surveillance sites. This network is able to detect and track orbiting objects, covering orbits ranging from LEO to GEO [3-5].

Space-Based Space Surveillance (SBSS) observation systems adopt a new approach expected to provide consistent improvements to the overall SD monitoring capability. [6]

The most important consideration on the SBSS system is that it can be considered as complementary to, and competitive with, the ground components. The SBSS system has several advantages comparing to the ground-based (GB) systems. First, it targets a specific portion of the debris surveillance work, which either cannot

be done from the ground, or which is at least very difficult and expensive from the ground. Moreover, other advantages are: the proximity to the small debris enabling in-situ measurements, the possibility of observing almost consecutively during 24 hours of survey (not to be interrupted by the daylight and no restrictions by weather), the absence of the atmospheric turbulence (no atmospheric seeing, diffraction limited design possible which means higher measurements accuracy) and of geographical and political restrictions. [7]

This paper presents the potential capabilities of a newly defined On-Board (OB) data processing system intended to a SBSS mission for Space Surveillance and Tracking (SST) based on a micro-satellite platform. SST is part of Space Situational Awareness and covers the detection, tracking and cataloguing of SD and satellites. The proposed system concept is based on ESA SST System Requirements and aims at fulfilling its core requirements.

The concept of the proposed system is based on a nearly continuous, high frequency, scanning of the sky, able to collect a big amount number of observations. Obviously, not all the collected data contain useful information, i.e. only when a potential SD is detected there is interest in saving OB and transmitting data to the ground (for further refined detection and processing). Thus, core of the system is the OB processing module able to pre-process and select data of interest for downlink.

2. IMAGE PROCESSING ALGORITHM

The OB processing module uses an algorithm for sources extraction that allows detecting, in automatic mode, the space debris in optical data using image-processing techniques.

The algorithm distinguishes both the feature present in the optical image (streaks and point-like objects) allowing its usage in both observation modes: sidereal tracking in which the star are point-like object and the space debris are streaks; object tracking in which the features are exchanged.

The extraction is performed using a single frame, furthermore the computation process does not use neither information about the image (observed zone and orbital regime of the observed object) neither the star catalogue to perform stars subtraction to detect streaks in the image.

The processing pipeline is composed of three main phases: the pre-processing step for noise reduction purposes obtained using non-linear digital filtering techniques; the segmentation and classification phase for the extraction of the connected components and the measurement of the shape properties; finally the astrometry phase computes the celestial coordinates of the detected objects. [8]

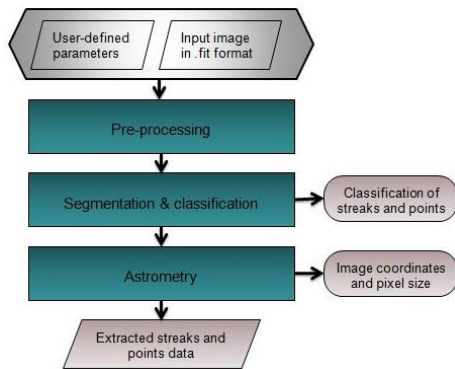


FIGURE 2-1 PIPELINE FOR SOURCE EXTRACTION

2.1. Pre-processing

In the preprocessing phase, we are interested to elaborate the input data in order to reduce the noise and prepare the image for the segmentation.

The first operation is the histogram stretching used to improve the contrast of an image by stretching the original dynamical range of intensity in a desired range.

Typically, SD image data are stored as 16-bit grayscale images in “fit” format; thus, in order to contextually reduce image dimension while preserving and highlighting image features, a remap of data in an 8-bit grayscale images is performed.

Once obtained such stretched image, a median filter is applied with the purpose of noise removal.

The median filter removes random impulse noise, it provides excellent noise-reduction capabilities, with considerably less blurring than linear smoothing filters of similar size. The median image is then analyzed to estimate the image background.

To detect faintest objects in the image it is necessary to compute accurate values of background level in the image.

In order to take account of variation of the background level in the image, a local analysis of the image is performed. The local statistics analysis estimates the background values in each mesh of a grid covering the whole image.

The obtained image is then subtracted from the median image to obtain a background-subtracted image, in which are present only foreground pixels.

2.2. Segmentation

The image segmentation includes all those operations that tend to partition an image into significant regions. The purpose of segmentation is to simplify the image information in order to make easy the features extraction. The first operation in image segmentation procedure is the image binarization.

The transformation in binary scale reduces the informative content of the image splitting the pixel in only two categories, foreground and background.

At this point, the binary image following two different segmentation procedure in order to enhance the researched features.

2.2.1. Segmentation for streaks detection purposes

In order to detect streaks, the distance transform is applied to the binary image.

The distance transform is performed by using a mask of 3-by-3 pixels, in which each point in the mask defines the distance to be

associated with a point in that particular position relative to the center of the mask.

The distance transformation result values are then normalized and the threshold to obtain the peaks value corresponding to foreground objects is defined.

Considering this normalized distance transformed image, a morphological dilatation filter is applied.

The application of a dilatation operator, actually bridges gaps and connects disjoint parts of the same object resulting from a threshold operation.

Then, aiming at measuring the streaks inclination angle, the Standard Hough transform is computed on the result image.

Finally, the results of the Standard Hough transform are used to apply a morphological opening filter. The morphological opening effects are preserve regions with shape similar to the structural element and deletes different ones. Using a linear structural element rotated of an angle α correspondent at the peaks of the Standard Hough transform, we preserve all the streak-like objects and delete all the other features.

2.2.2. Segmentation for points detection purposes

The first operation to detect the point-like objects in the image is the application of a convolution filter. A square kernel of 3-by-3 pixels scans the image and replaces all image pixels under the central point of the kernel with the value 1 if the sum of the image pixel under the kernel is higher than a threshold. A higher threshold value allows deleting single points as hot pixels or cosmic ray.

The next operation is the morphological opening. Using a circular structural element, we obtain the removal of linear object as streaks or noise effect preserving the point-like objects.

Finally, to ensure the deletion of all streak-like objects from the image, a subtraction operation is performed between the morphological opening image and the binary image obtained in the segmentation for streaks detection purpose image.

With the end of the segmentation phase, we presume that the obtained images allows identifying clearly the object contours and then classify them as stars or streaks.

2.3. Classification

The classification phase starts with the identification of the objects contours. A contour is a list of points that represent a curve in an image. We assume that a pixel is a contour pixel if it is a white pixel and if it has at least one adjacent black pixel in his surroundings. Finally, to obtain all pixels inside the contour, the Ray-casting algorithm is applied.

Once terminated this identification phase, all the detected objects are measured to distinguish if they are stars or streaks.

To compute the object barycenter and elongation we use the formulation of the image Moment:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y) \quad (1)$$

being x and y the coordinate of the pixel belonging at an object and $I(x, y)$ the pixel intensity. By this definition we obtain that the Moment of zero order is the area of the object express in pixel, while the ratio of Moment of first order with the Moment of zero order are the coordinate of the object’s barycenter.

Other descriptors of the object are the Central Moments

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y) \quad (2)$$

and the centered Central Moments

$$\mu'_{pq} = \frac{\mu_{pq}}{\mu_{00}} \quad (3)$$

The centered Central Moments are used to describe the object as an elliptical shape, obtained the semi-major a and semi-minor b axis of the ellipse.

With these measures explained above, we classify all the detected objects as point-like or streak-like objects.

A first selection of the objects is performed taking into account the object's dimension. A detected object is rejected if the zero order Moment is lower than a threshold σ . For the threshold σ , a value of 5 has been selected, in this manner all the false positive detection objects resulting from noise or artifact are discarded.

The remaining objects are then studied to classify them as point-like or streak-like objects. The study is based on the analysis of the object's semi-major and semi-minor axis.

An object is classified as point-like object if the inequality $a/b < \rho$ is satisfied. Ideally, the ρ parameter should be equal to one; but considering the spreading of the photons, the effect of the image noise, and the artifact due to the image processing it is selected a value of $\rho = 1.6$.

On the contrary, an object is classified as streak if the inequality $a/b < \delta$ is satisfied being $\delta > \rho$. The δ parameter is function of the observed object orbit, exposure time, optical and camera features. Its value has been selected after a study of the usual dimension and shape of the space image features taking in different orbital debris, from LEO to GEO, in order to be suitable for every image type.

3. GPGPU

The use of General Purpose computing on Graphics Processing Units (GPGPU) is a technological choice aimed at increasing the computational performance of scientific and engineering applications for large-scale parallel processing applications [9]. In 2006, the NVIDIA Company with the Compute Unified Device Architecture (CUDA) has released the Application Programming Interface (API) and Software Development Kit (SDK) with the intent to simplify the accessibility and the use of the GPU. The CUDA-C language is fully integrable with the C++ code and its APIs are quite similar to those of the C language.

3.1. GPU program

From a hardware standpoint, the GPU architecture includes a host, which is a traditional CPU architecture, and device, which is a massively parallel processor (the GPU).

The host drives the computational process by a CUDA program. The CUDA program is a heterogeneous code consisting of many parts having phases that can execute both on the host and device, thus having a unique source code containing both host and device code. The host code is written in C++, while the device code is written in CUDA-C code, which is an extended version of the C language with special keywords for labelling data-parallel kernels and their associated data structures. Usually a CUDA program is composed at least of these phases: read input data, copy input data from host memory to device memory, process the data on the GPU by parallel kernel and finally copy result data from device memory to host memory

A kernel is a function written in CUDA-C language that executes parallel code, it can run only on NVIDIA's GPU. The kernel is executed by each GPU's thread; threads are identified by a unique

ID, enabling the programmer to address different parts of GPU memory relative to the thread ID.

In the synchronous computation only one thread is responsible of the execution of a function on the image pixels, then the function has to perform the mathematical computation and the selection of the image (for example the application of a moving window filter like the median filter). In the parallel execution, many threads are launched at the same time, each one performing the mathematical computation for one pixel or an image region. This implies that the kernel must be written in order to execute independently of the selected memory area and invariants of the execution order.

3.2. CUDA threads

CUDA organizes threads in a Scalable Programming Model: the GPU's threads are grouped into blocks (mono-/bi-/tri-dimensional) and identified by means of a thread index called *threadIdx*; in turn blocks are grouped in a grid (mono-/bi-dimensional) and identified with a block index called *blockIdx*. The GPU has a finite number of threads per block and a finite number of blocks per grid, and therefore has a limit to the amount of parallel execution. If the number of blocks exceeds the max limit, then the GPU sequentially processes the maximum possible number of blocks, therefore a kernel can be considered as executed in parallel manner in function of the hardware limit.

4. RESULTS

The proposed method was tested and validated on a set of data containing both imaging satellites from different orbit ranges and multiple observation modes (i.e. sidereal and object tracking).

Figure 4-1A shows the detection results for the GEO image in which two objects with different orbits are present.

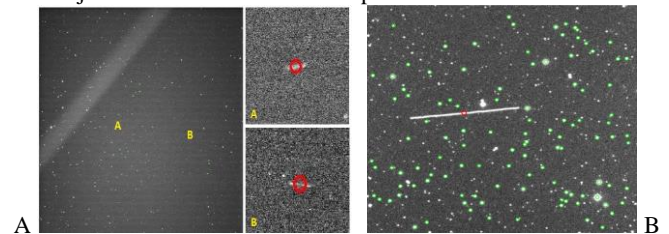


FIGURE 4-1 GEO AND LEO IMAGE IN SIDEREAL TRACKING

Figure 4-1B shows a space debris in LEO orbit. Detected streaks are marked in red while point-like objects are marked in green. In this case, of images taken with sidereal tracking, they correspond respectively to space debris and stars.

An example of object tracking image is shown in Figure 4-2. In this case, unlike the previous one, the detected streaks and point-like object correspond to stars and space debris respectively.

The total number of processed images is 400, 547 objects were checked by visual inspection in 356 images.

The pipeline detects correctly 469 objects in 308 images with a success ratio of 85%.

4.1. Timing and speedup results

In this section, we analyse the performance of the proposed speed up process.

All the presented results were tested on a NVIDIA Jetson Tegra K1 (TK1), a quad-core 2.3 GHz ARM Cortex-A15 CPU processor with

2GB of RAM with a NVIDIA Kepler GPU of 192 cores capable of over 300 GFLOP/s of 32-bit floating-point computation compatible for on-board processing.

The processed images have different dimensions, ranging from 1024x1024 to 4096x4096 pixel. This characteristic is of fundamental importance in the speedup analysis. To take into account the variation of the speedup with the image's dimension, the computational time has been evaluated using images of several dimension of the same sky zone; the set of analysed dimensions is: 128x128, 256x256, 512x512, 1024x1024, 2048x4048 and 4096x4096 pixel.

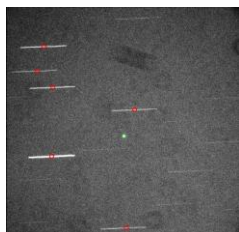


FIGURE 4-2 GEO IMAGE IN OBJECT TRACKING

All the measured times are the mean times of ten executions of the algorithm, this allow to avoid corrupted time due to others operating system process that run in parallel during the execution of the algorithm.

Figure 4-3 depicts the execution time of the algorithm in serial using a single CPU thread (blue line) and parallel using all the GPU thread (green line) mode; in which with serial mode we intend the code without parallelization that run only on CPU, while with parallel code the version with the use of CUDA that run on CPU and GPU.

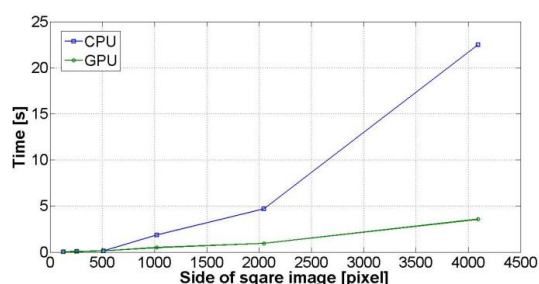


FIGURE 4-3 EXECUTION TIME CPU VS. GPU

In Figure 4-4 is plotted the trend of the speedup, these values indicates the time reduction with respect to the execution time of the serial version of the algorithm $S = t_{CPU}/t_{GPU}$.

The values of the speedup show important results.

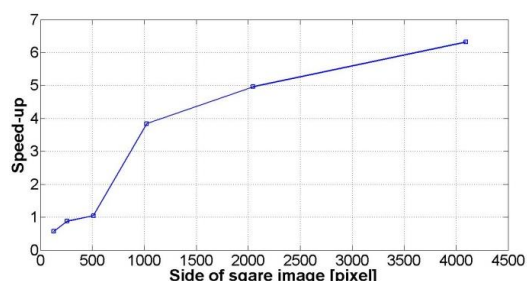


FIGURE 4-4 SPEEDUP

For image with dimension up to 512x512 pixel, the time saved in the parallel processing is not great enough to warrant the time taken for transferring the data between computer memory and GPU

memory. For this reason it is preferable analyse image with little dimension on CPU with serial code.

The potentiality of the GPU are evident with image of dimensions higher then 512x512 pixel.

For these dimensions, the time saved in the parallel processing is very high enabling the possibility to obtain a relevant time reduction. The best speedup value has been measured for images of 4096x4096 pixel; the execution time has been 22.48 seconds for the CPU version of the pipeline and 3.55 seconds for the GPU version with a speedup factor of 6.3x.

5. CONCLUSIONS

A pipeline based on GPU technique for sources extraction in automated image analysis for space surveillance applications has been proposed and its speedup and the efficiency discussed. The pipeline detects space debris without any a priori information and it is based on the analysis of a single image. The pipeline is able to detect space debris in both the sidereal and object tracking modes. The GPU approach was introduced in order to reach near real-time performance. The mean execution time for image of 4096x4096 is 3.55 seconds with a speedup of 6,3x with respect to the CPU-based system. The pipeline was tested on 400 images with a success ratio of 85%. The pipeline develop with Jetson TK1 embedded platform prove its use as autonomous on-board objects detection able to optimising the downlink bandwidth usage.

6. REFERENCES

- [1] Klinkrad H. Et al, "Space Debris Activities in Europe," *4th European Conference on Space Debris*, 18-20 April 2005, ESA/ESOC, Darmstadt, Germany, p.25.
- [2] Kessler D. J. Et al, "The Kessler Syndrome: Implication to Future Space operations," *33rd Annual AAS Guidance and Control Conference*, 6-10 February 2010, Breckenridge, Colorado
- [3] Piergentili F. et al. "EQUO: an Equatorial Observatory to improve the Italian space surveillance capability," *66th IAC*, Jerusalem, Israel, 2015, 12 – 16 October.
- [4] Diprima F. et al. "Lessons learned in automatic operation of observatories for space debris observation," *67th IAC*, Guadalajara, Messico, 2016, 26 – 30 September.
- [5] T. Cardona, et al. "The Automation of the Equo On-Ground Observatory at Broglio Space Center for Space Surveillance," *67th IAC*, Guadalajara, Mexico, 2016.
- [6] J. Utzmann et al. "SBSS Demonstrator: A Space-Based Telescope for Space Surveillance and Tracking," *7th IAASS Conference*, Friedrichshafen, Germany, 2015.
- [7] J. Utzmann, et al. "Optical In-Situ Monitor Breadboard System," *7th European Conference on Space Debris*, Darmstadt, Germany, 2017.
- [8] Nixon M. S., Aguado A. S., *Feature Extraction and Image Processing*, Newnes, 2002.
- [9] NVIDIA Corporation, [Online], CUDA Toolkit Documentation v8.0.61, Available from: <http://docs.nvidia.com/cuda/#>, 2017

YOUNG TREE IDENTIFICATION USING MACHINE LEARNING ON SENTINEL-1 DATA

S. Daniel¹, J. Klein¹, J. Petrat², Y. Lee², L. Brown²

¹Capgemini, France

²Capgemini, UK

ABSTRACT

The Forestry Commission (FC) spends 20-40% of its budget monitoring, administering, and pursuing the establishment of young tree stock by aerial photography and on-site inspections. Capgemini proposes to UK Environmental Agency an improved monitoring system combining the used of Space data processing and Big Data technologies expertises. Indeed, the innovative solution presented here proposes to engineer features from SAR data along with auxiliary as inputs and machine learning models from Data Science field. The inputs used are (1) Sentinel 1 data products which provide a high frequently measurements, (2) in-situ data form The National Forest Estate (NFE) and the National Forest Inventory (NFI) providing shapefiles, (3) meteorological data coming from UK MetOffice. The applied methodology is a SAR-machine Learning framework relying on pre-processing with Sentinel-1 toolbox, reprojection and multi temporal filtering approach, exploratory analysis, feature engineering and statistic machine learning to obtain actionable insights accessible through a web application offering visual analytics. The promising results show that a predictive saving in monitoring operations up to 20%.

Index Terms— Sentinel-1, SAR, Big Data, Machine Learning, Cloud computing, concrete use case

1. INTRODUCTION

The Forestry Commission (FC) spends 20-40% of its budget monitoring, administering, and pursuing the establishment of young tree stock. Private landowners can request subsidies from the EU to plant and grow new forests on their land. They can also apply for subsidies to restock areas of trees that have been felled. In case of a failed scheme, like landowners not planting any trees, such subsidies can be claimed back by the government within ten years after the agreement. The National Forest Registry classifies land that has received subsidies as “assumed woodland”. Monitoring the status of assumed woodland starts with the identification of young trees, defined to be planted no more than seven years before the current date. FC employs two methods of monitoring assumed woodland:

- Aerial Photography (AP):

Since this process involves a lot of manual work and since AP is updated very infrequently, it is difficult to respond to violations by landowners in time.

- On-site inspections:

FC can send staff to sites with assumed woodland to inspect whether young trees have been planted. Given the sheer amount of assumed woodland and the limited resources of FC, less than 10% of all sites can be monitored every year.

The solution considered here is to use Sentinel 1 Synthetic Aperture Radar (SAR) data to identify young trees. As opposed to aerial photography, the Sentinel 1 constellation takes SAR measurements frequently: approximately every five days. Analysing this data will enable a much more timely response to violations of

agreements and a close-knit monitoring of assumed woodland. Furthermore, SAR backscatter information can be used to distinguish clear-fell or unused land from regions with newly planted trees by tracking the temporal change in height of the land. By deploying the proposed solution and relying more heavily on SAR remote sensing, this operational cost of £30m to £60m can be expected to be reduced by up to 20%, which amounts to potential savings of up to £12m p.a.

2. DATA

The solution proposed here is to engineer features from SAR data along with auxiliary as inputs for machine learning models. The technology tools used to carry out the investigation were open source programs ran on Amazon Web Services, licensed EC2 instances and Elastic MapReduce (EMR).

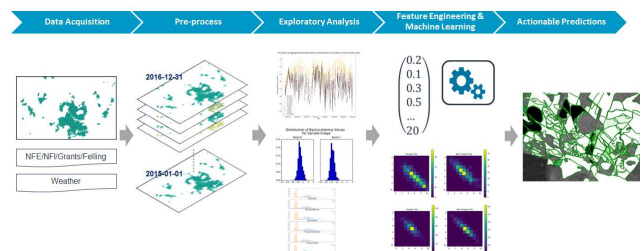


Figure 1: Overview of the framework

2.1. SAR data

For the purpose of this research, 2 years of SAR data (2015-2016) have been downloaded. Then, the related data set has been pre-processed with Sentinel 1 toolbox from SNAP, a free software provided by European Spatial Agency. The preprocessing steps include: Calibration and Terrain Correction for a ground spacing pixels fixed at 10 m square. In case of overlaid of images across time, a re-projection step is performed. It is parallelised using Spark running against S3. Besides, we are in the case of a time series study. In order to filter the speckle, the preferred chosen approach is to remove it with a Multi temporal filtering [1][2] instead of the Lee filtering method.

2.2. Ground measurement

To obtain a categorisation of the various woodlands in regions of interest, two publicly available shape files from the FC are used:

- National Forest Estate (NFE): The NFE contains the FC sub-compartment database and gives information for recording, monitoring, analysis and reporting across the entire FC estate.

- National Forest Inventory (NFI): The NFI includes interpreted forest types for all woodlands over 0.5 ha and interpreted open area information for areas over 0.5 ha surrounded by woodlands.

2.3. Weather data

In order to capture seasonality and environmental influences on SAR backscatter, weather data was collected for all regions of interest. The Darksky API provides hourly data for humidity, pressure, temperature and wind speed. Additionally, monthly rainfall data was obtained from the Met Office.

Every SAR image is split to cover individual polygons, and for every such combination of SAR image and polygon, a record in a data frame storing all combined data is created.

3. FEATURE ENGINEERING AND MODEL

The outputs of the model are probability values for the polygon to contain sufficient levels of young trees. If the probability is above a threshold, the decision will be there are young trees in that polygon. The threshold has been set to 0.5. Features are representations of the available data which together with a model will determine the young trees presence. The feature engineering is designed to be an iterative process. Within the PySpark framework, a script has been developed retrieving a list with features and sequentially applies them to a dataset. The related output is a ‘feature data frame’ which contains all the computed features and a unique identifier for every SAR image. The features created together with the ground truth are used to train the model on the training and validation data and this latter is applied to the test data. An analysis of the results on test data reveals if a feature improves the accuracy of the predictions. Features with a positive impact are selected and refined.

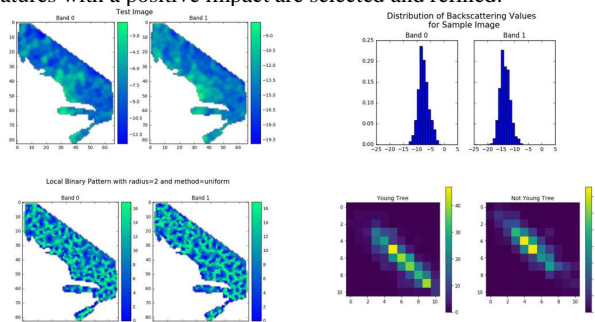


Figure 2: Actual features

As there are significantly more polygons in categories different from young trees, all accuracies are reported on a subsampled dataset with equal number of elements in each predicted category.

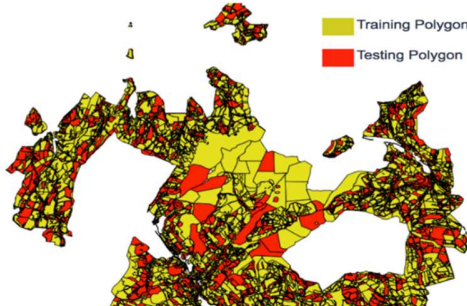


Figure 3: Training and testing polygons over UK

The figure 3 shows the test polygons in red (30%). The polygons in yellow are split into training and validation polygons with a 70-30% split. According to our data, the machine learning algorithm

XGBoost with decision trees [3] has the best performance of all models tested.

4. RESULTS

Due to the complexity of the problem, it has been selected a simple region of interest for the first trial, Kielder. Then, further regions would be selected for testing the scalability aspect: East Anglia and Cambrian Mountains. East Anglia has a flat topography and known for large quantities of weed in the forest regions. Together, Kielder, East Anglia and the Cambrian Mountains represent a great share of woodland in UK. Two experiments that best showcase the practical accuracy of the model were realised: “young trees vs. non-woodland” and “young trees vs. grassland”. The latter experiment is considered to be a slightly higher difficulty given the non-woodland category can contain easily distinguishable roads, rocks, buildings, etc. The model achieves to a classification score of 85% for both experiments for Kielder region. It performs particularly well when applied to trees in the year they are planted, which is likely due to de-vegetation and cultivation. Results on East Anglia are good also with respectively classification scores of 80% and 81%. These scores are very promising and may already be sufficiently high for many applications.

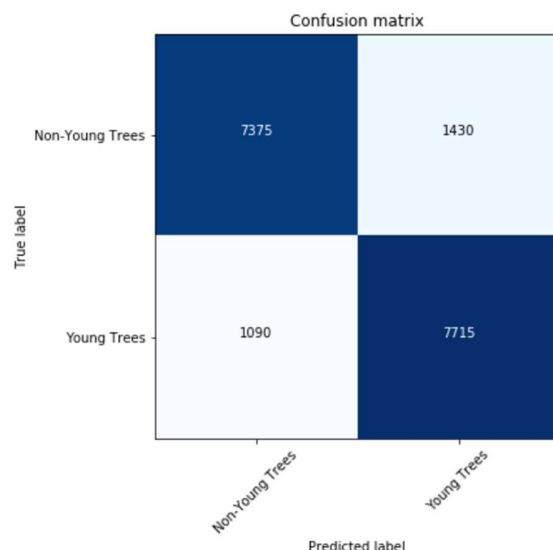


Figure 4: Confusion matrix

5. SCALABILITY

The results obtained demonstrate that the SAR-Machine Learning framework is able to produce a machine learning model that can identify young trees with a high accuracy. It can be considered scalable if it meets the two requirements above:

- Parallelisation:

To evaluate if the underlying infrastructure and the computations in each processing step can be processed in parallel, we chose to run experiments on EMR with varying cluster sizes of 2-9 worker nodes. We focused on investigating the scalability of the five most time-consuming processing steps on the Kielder region. All of these were successfully executed on EMR. Moreover, it could be observed that computations were evenly distributed over executor nodes, providing more evidence of the horizontal scalability. The success

of the experiments strongly suggests that the SAR-Machine Learning framework is horizontally scalable. Nevertheless, scaling has been only tested on a subset of UK; further large-scale tests would be required to obtain estimates for the total processing time.

- Applicability to other regions of Great Britain:

Accurate predictions in all three regions would provide strong evidence for the global capabilities of the machine learning model. Despite the good results obtained for Kielder and East Anglia, the settings of our DEM used for terrain correction didn't allow to perform the analysis on Cambria Mountains. It was due to high slope variation, as the terrain simply cannot be located accurately.

6. FUTURES

There is an opportunity for improving the existing solution and expanding it to work for even wider regions in UK:

- Enhancement of existing features & development of new ones:

There are many variants of the existing features that could be worked on in addition to the development of new features. It would improve the model accuracy for existing and additional tests.

- Trial of Defra DEM

The existing DEM used for terrain correction has a low resolution of 90 m x 90 m. With a higher resolution DEM, it may be feasible to work on mountainous regions and also to improve the accuracy of our existing models by reducing noise.

- Industrialised architecture:

We have demonstrated some of the components are scalable and highlighted challenges in the standard EMR set-up. We then would recommend the definition of a scalable architecture.

7. CONCLUSION

This work has demonstrated that a significant proportion of regions can be modelled with sufficiently high accuracies using basic feature engineering. However, it is not yet feasible to build models for mountainous terrains, potentially due to the distortion of DEM. Hence, this issue could probably be resolved by using a higher resolution DEFRA.

Another recommendation is the definition of an industrialised and scalable architecture of the framework. In the long term vision, the young tree identification model may be implemented as a tool to help business analysts with automatic recommended decisions derived from model predictions, as well as give business intelligence and insight for data mining and big data studies.

8. REFERENCES

- [1] J. Bruniquel, A. Lopes, "Multi-variate optimal speckle reduction in SAR imagery", *Int. J. Remote Sens.*, vol. 18, no 3, pp. 603-627, February 1997
- [2] S. Quegan, J. J. Yu, "Filtering of multichannel SAR images", *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no 11, pp. 2373-2379, November 2001
- [3] S. Tufféry, "Data Mining and Statistics for Decision Making", John Wiley & Sons, 2011

THE SIX FACES OF THE DATA CUBE

*Peter Strobl¹, Peter Baumann², Adam Lewis³, Zoltan Szantoi^{1,4}, Brian Killough⁵,
Matthew Purs³, Max Craglia¹, Stefano Nativi^{1,6}, Alex Held⁷, Trevor Dhu³*

¹European Commission, Joint Research Centre, Ispra, Italy; ²Jacobs University, Bremen, Germany;

³Geoscience Australia, Canberra, Australia; ⁴Stellenbosch University, South Africa;

⁵NASA Langley Research Center, Hampton, United States; ⁶National Research Council of Italy, Rome, Italy;

⁷Commonwealth Scientific Industrial Research Organisation (CSIRO), Canberra, Australia

ABSTRACT

This paper provides a structure to the recently intensified discussion around ‘data cubes’ as a means to facilitate management and analysis of very large volumes of structured geospatial data. The goal is to arrive to a widely agreed and harmonised definition of a ‘data cube’. To this end, we propose an approach that deconstructs the ‘data cube’ concept into distinct aspects. We have identified six such aspects, which we refer to as the *6 faces* of the data cube. More than a pleasing analogy, these *6 faces* are fairly independent, and hence ‘orthogonal’ domains. They should allow breaking down the description and handling of data cubes into meaningful and manageable ‘parts’, which however, only if seen holistically, make it possible to harness the full potential of this multidisciplinary infrastructure.

Index Terms— data cube, structured data, data infrastructure, geospatial data, big data, standardisation, WCS, CIS, INSPIRE, OGC, ISO

1. INTRODUCTION

The term data cube originally was used in Online Analytical Processing (OLAP) of business and statistics data; technically speaking, such a data cube represents a multi-dimensional array together with metadata describing the semantics of axes, coordinates, and cells. More recently, data cubes have emerged in a geospatial context [1,2] as an approach to the management and analysis of these large and rapidly growing datasets. While the terminology ‘data cube’ was used as early as in the 1980’s when the first imaging spectrometers produced ‘hyperspectral data cubes’, technology was not ready for efficiently storing and serving data cubes. Geospatial data cubes typically are densely populated, whereas OLAP data cubes typically are sparse. A generic requirement remains that data can only be organised as a ‘cube’ if they have inherent attributes (usually referred to as coordinates) according to which they can be ordered. A data cube may have horizontal and vertical spatial axes, temporal axes, or any other application-dependent dimensions. For geospatial data cubes, at least one of those should be non-spatial.

2. DEFINING THE CUBE

Similar to the term ‘big data’, for which no consistent definition has yet emerged, we find the notion of ‘data cube’ still varying across the literature and often dependent on the context in which it is used. Whilst ‘big data’ is an expression at a high and abstract enough level so that a certain room for interpretation is not problematic, discussions around data cubes will suffer, unless further structure is provided to this evolving concept.

For us, a Geospatial Data Cube (GDC) is based on regularly and irregularly gridded, spatial and/or temporal data with *n dimensions* (or *axes*) and characterised by the presence of the *6 faces* that we explore in this paper. As such, it complements the conceptual view of the ‘Datacube Manifesto’ [4] with a holistic system view, whose aim is to raise awareness for all necessary aspects of such an infrastructure.

3. DISSECTING THE CUBE

The purpose of a GDC is to allow ingestion, storage, provision, and analysis of structured geospatial data for which it has to cover several technical aspects, which we call, *faces*. Individually each face is a well-established domain within data sciences, allowing the respective experts to enter the discussion at the right end. However, as an infrastructure a data cube can unfold to its full potential only if all the following ‘faces’ are comprehensively covered and well-orchestrated.

3.1. Parameter Model

The semantics of a cube cell value is described by a parameter model which allows understanding the information stored in each thematic layer of the cube. This includes the parameterisation of the property and its quality, as well as the associated metadata that are necessary for the analysis. The Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) Common Data Model (CDM) [3] defines important elements of parameter models. Well-documented implementations of such models for various themes, such as terrain elevation [5], are given in the INSPIRE data specifications. However, incorporating data

describing the same parameter data (i.e. radiance imagery) but from various origin into a geospatial data cube remains a challenge even in cases where such models are applied, due to the differences among collecting sensors, imagery processing chains and algorithms used. Thus, such geospatial (raster) data need to be either pre-processed with approved algorithms or, rather should be directly produced by the corresponding instrument owner such that they fit into the data cube structure. The latter, and preferred option is being advocated and endorsed by the Committee on Earth Observation Satellites (CEOS). Such data, called “Analysis Ready Data” or ARD, would come from CEOS’ member space agencies and fulfil a minimum set of criteria, like consistent parameter models and approved algorithms, thus largely facilitating the compilation of data cubes and data exchange among them. Direct or automatic multi sensor data fusion however calls also for harmonised sensor characteristics such as spectral band definition and availability and consistency of ancillary data like Digital Elevation Models.

3.2. Data Representation

Data representation is the way in which a parameter is discretised and semantically encoded along the different axes or dimensions of the cube such as space, time, and thematic properties. A given parameter might be represented in different ways and the same representation scheme might be used for different parameters. Depending on the representation type a specific set of metadata needs to be supplied including e.g. range, interval, scale, precision, or reference. The OGC SWE-CDM contains a comprehensive overview of representation types [3].

Discretisation in the spatial domain is highly familiar in the form of gridding [6]. ISO and OGC today base most of their grid definitions on the EPSG catalogue of projections, which either limits respective grids to regional coverage or induces considerable spatial distortion. An example for a common (quasi) global spatial grid system is the WMTS, which in fact is a mixture of projection, grid definition and tiling schema. A relatively new concept is promoted by the recent OGC standard for Discrete Global Grid Systems (DGGS), which aim at overcoming limitations of planar projections by defining hierarchical grids directly on the ellipsoid.

In other areas standards are often still missing, and the representation of observation-level metadata such as measurement quality and uncertainty is in its infancy.

3.3. Data Organisation

The cell values generated by the discretisation of the parameter need to be physically arranged and stored in a machine-readable way. This encompasses issues like file formats, file systems, and database structures. OGC CIS [6] - which is also adopted as ISO 19123-2 - establishes how representation can be based on ASCII (such as GML, JSON,

or RDF), binary (such as GeoTIFF or NetCDF), or a mix of both embedded in some “container format” (such as zip or GeoPackage). Furthermore, the data cubes representing “Big Data” typically require data to be partitioned (also called tiling), and they need to be amenable to streaming (mainly in case of timeseries); both is included in the current version CIS 1.1. Furtado [7] performed a general analysis of multi-dimensional partitioning.

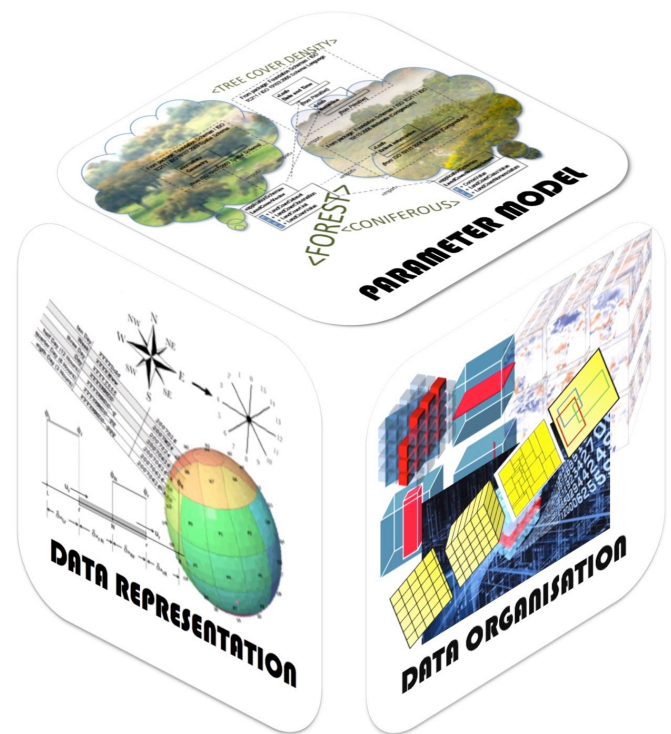


Fig. 1 The Data oriented Faces of the Geospatial Data Cube.

3.4. Infrastructure

The data storage units must be hosted by an IT infrastructure or ‘hardware’ that also allows their handling. This could be a centralised or distributed setup of storage and processing devices. Rapid data access and transfer between storage and processing instances are important criteria [2], particularly for very large spatio-temporal datasets.

Amount and increase of geospatial data require significant financial and logistical investments to offer competitive services for attracting and retaining users. Among the many supercomputing facilities, which over the last years have started offering geospatial data and services are industrial initiatives such as the Google Earth Engine [8] or Amazon Web Services. Others are publicly funded and operated such as the Australian Geospatial Data Cube [2], the Technical University of Vienna’s Earth Observation Data Centre (EODC) [9] or the JRC Earth Observation Data Processing Platform (JEODPP) at European Commission’s (EC) Joint Research Centre [10]. In the frame of the Copernicus program the EC is about to fund various consortia uniting

public and private entities to serve as ‘Data Information and Access Systems’ (DIAS) [11,12].

While all these initiatives also show commitments covering other aspects of data cubes, their main investments seem to be directed towards the IT infrastructure. However, the success of these investments will largely depend on the functionality of these infrastructures for which they must also duly cover the other faces described here.



Fig. 2 The functionality oriented faces of the Geospatial Data Cube.

3.5. Access and Analysis

Within the infrastructure a wide range of functionalities must be implemented through software to access, manipulate and analyse the stored data (and metadata) and to ingest new products into the data cube. These functionalities must be documented and made available to users by means of APIs and other interactive interfaces (GUIs). Between the User API (front-end) and the file manipulation routines (back-end) one or several layers of software are imaginable.

One of these layers could consist of common GIS tools (e.g. QGIS, ArcGIS), and OGC Web Coverage Services (WCS) can be used to connect these within the data cube. A most recent example of an API and GUI has been demonstrated by the CEOS Open Data Cube initiative (<http://tinyurl.com/datacubeui>).

An existing standard defining a GDC analytics language is the OGC Web Coverage Processing Service (WCPS) [13].

Additional recent attempts to establish such languages are made by OPeNDAP, Google Earth Engine [8] and others.

As substantial processing is being shifted to the data cube host, anticipative cost estimation as well as access rights and security will also be of high concern when it comes to granting access to data and to analysis power. Given the size of data cubes it will often not be sufficient to give a binary answer on the whole cube, but guard particular regions, collections, etc. separately. Costs for accessing, processing, and transferring data should be determined prior to execution so that the host can decide about admissibility, and maybe users can be warned or disproportionate request rejected.

3.6. Interoperability

Interoperability and scalable fusion of spatial information across different data cubes is crucial and highly dependent on the use of robust international standards governing the access and transfer protocols for communication between client and server as well as among different servers.

ISO 19123 (which is identical to OGC Abstract Topic 6) defines an abstract data cube model as part of the coverage concept; however, due to its level of abstraction it is not yet interoperable. Its sister standard, OGC CIS 1.1 / ISO 19123-2, establishes concrete encodings which allow re-encoding of coverages from one format into another so that a well-defined, format-independent data cube exchange is possible, though at the cost of additional interpolation and resampling.

The corresponding service model is provided by the OGC Web Coverage Service (WCS) [13], which has been adopted by INSPIRE and is on the adoption plan of ISO. A large, growing number of open-source and proprietary implementations support WCS so that interoperable access to data cubes is possible through a wide range of tools today, including map navigation (like OpenLayers, Leaflet), Web GIS (like QGIS, ArcGIS), visualization (like NASA WorldWind, Cesium), and analytics (like python and R) - see the examples in the Jupyter notebook at [14]. This allows users to remain in the comfort zone of their tools while accessing data cubes stored in rasdaman, GeoServer, MapServer, ArcGIS, and other WCS-enabled engines.

Further, the Web Coverage Processing Service (WCPS) geodatacube analytics language standard provides a means for “shipping code to data” in an unambiguous, semantically well-defined manner [13].

Since 2012, the intercontinental EarthServer initiative (<http://www.earthserver.eu>) is establishing agile datacube analytics on 3D x/y/t image timeseries and 4D x/y/z/t weather data, based on the rasdaman Array Database System (<http://www.rasdaman.org>). The largest installation, EO Data Service (www.eodataservice.org), recently has passed the 1 Petabyte frontier; ECMWF in EarthServer is working on unleashing its 220 PB climate archive. Currently many more stakeholders such as the Committee on Earth

Observation Satellites (CEOS) and W3C have started working on data cubes. Consistency among these and the established OGC / ISO / INSPIRE standards will be a key to success. Barriers to interoperability, on the other hand, will inevitably lead to silo effects undermining the multidisciplinary concept and potential of data cubes.

4. OUTLOOK

The future success of (geospatial) data cubes will certainly not depend on the existence of a widely-agreed definition alone. But it is likely that a well-structured discussion and a widespread agreement on key features of data cubes will enable a much faster convergence, increased interoperability and more rapid progress at global level.

Valuable technology contributions can be expected from the field of Array Databases, which is working on flexible, scalable query services on massive arrays, backed by the existing OGC Web Coverage Processing Service (WCPS) [13] and the forthcoming ISO Array SQL [16] standards.

However, users should not need to learn new languages each time they work on another platform, but be able to use their own existing tools and scripts (e.g., python and R for analysis), which can be coupled through the abovementioned languages as hidden, standards-based client/server APIs.

Ultimately, the efforts should go beyond just the exchange of data, but move us towards compatibility and consistency of the available information and of the way it can be accessed and analysed.

5. REFERENCES

- [1] Salehi, M., Bédard, Y., Mostafavi, M., Brodeur, J., 2007, "From transactional spatial databases integrity constraints to spatial data cubes integrity constraints", Proc. of the 5th International Symposium on Spatial Data Quality.
- [2] Lewis, A., et al., 2017, "The Australian Geoscience Data Cube — Foundations and lessons learned", Remote Sensing of Environment, <http://dx.doi.org/10.1016/j.rse.2017.03.015>
- [3] Robin, A. (Ed.), 2011, *SWE CDM Encoding Standard*, OGC, <http://www.opengeospatial.org/standards/swecommon>
- [4] Baumann P., 2017, "The Datacube Manifesto", <http://www.earthserver.eu/tech/datacube-manifesto>
- [5] *INSPIRE Data Specification on Elevation – Tech. Guidelines* <https://inspire.ec.europa.eu/file/1530/download?token=pq85sbLG>
- [6] Baumann, P., Hirschorn, E., Maso, J., 2017, *Coverage Implementation Schema, version 1.1*, OGC, https://portal.opengeospatial.org/files/?artifact_id=48553
- [7] Furtado, P. et al, 1999, "Storage of Multidimensional Arrays based on Arbitrary Tiling.", ICDE'99, Sydney, Australia
- [8] Gorelick, N., et al., Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment(2016), <http://dx.doi.org/10.1016/j.rse.2017.06.031>
- [9] Wagner, W., 2015, Big Data infrastructures for processing Sentinel data, in Photogrammetric Week 2015, Dieter Fritsch (Ed.), Wichmann/VDE, Berlin Offenbach, 93-104
- [10] Soille, P., et al, 2017, The JRC Earth Observation Data and Processing Platform, Big Data from Space BiDS'17, this issue
- [11] <http://copernicus.eu/news/upcoming-copernicus-data-and-information-access-services-dias>
- [12] Schick, M., 2017, EUMETSAT, ECMWF & MERCATOR OCEAN partners DIAS
- [13] Baumann, P., 2009, *Web Coverage Processing Service (WCPS) Language Interface Standard*, OGC, <http://www.opengeospatial.org/standards/wcps>
- [14] Baumann, P., 2012, "OGC Web Coverage Service (WCS) Core", OGC, <https://portal.opengeospatial.org/files/09-110r4>
- [15] Clements, O., et al, 2017, "Improving access to big data through OGC standard interfaces", <https://nbviewer.jupyter.org/github/earthserver-eu/INSPIRE-notebooks/blob/master/index.ipynb>
- [16] Misev, D., et al., 2015, "A Database Language More Suitable for the Earth System Sciences". In G. Lohmann et al (eds.): *Towards an Interdisciplinary Approach in Earth System Science* Springer 2015, doi:10.1007/978-3-319-13865-7

DIGITAL EARTH AUSTRALIA – UNLOCKING INNOVATION AND CAPABILITY

Trevor Dhu, David Gavin, David Hudson, Trent Kershaw, Adam Lewis, Leo Lymburner, Norman Mueller, Simon Oliver, Jonathon Ross, Andreia Siqueira, Medhavy Thankkapan¹

¹Geoscience Australia, Jerrabomberra Ave & Hindmarsh Drive, Symonston ACT 2609, Australia, email: Earth.Observation@ga.gov.au

ABSTRACT

The Australian Government is investing to establish an operational large-scale analysis platform for satellite imagery and other Earth observations. From sustainably managing the environment to developing resources and optimizing our agricultural potential, Australia must overcome a number of challenges to meet the needs of our growing population. Digital Earth Australia (DEA) will deliver a unique capability to process, interrogate, and present Earth observation satellite data in response to these issues. It will track changes across Australia in unprecedented detail, identifying soil and coastal erosion, crop growth, water quality, and changes to cities and regions. DEA is based on the Open Data Cube [1], an open source technology and community that is focused on increasing the value and impact of global Earth observation data.

Index Terms— Data Cube, Digital Earth Australia, Big Data, Spatial Industry

1. INTRODUCTION

Australia's surface has been continually imaged by satellites for decades, recording a wide range of information about Australia's land and water resources. Historically, this data has been warehoused in unreliable and difficult to access government stores, where its potential is wasted.

DEA translates over 30 years of Earth observation satellite imagery into information and insights about the changing Australian landscape and coastline, providing a groundbreaking approach to organizing, analyzing, and storing vast quantities of data. It provides access to businesses, researchers, and governments to monitor and track these changes over time. To fully realize the benefits of DEA once operational, the platform and products will be open and freely available to any user. DEA will provide governments, individuals, and businesses with reliable, standardized, and easily accessible products and services; which will deliver new capabilities to increase efficiency, bolster profit, and create jobs. This will revolutionize land planning, agriculture, mining, environment analysis, and research.

2. CONCEPT

DEA is a series of data structures and tools which organize and enable the analysis of large Earth observation satellite data collections. A key element of DEA is the calibration

and standardization of the data. This increases the value which can be derived from Earth observation and other sources of large datasets, as it allows for the rapid development of information products to enable informed decision making across government and private industry. In the past, satellite imagery and other geospatial datasets were downloaded, analyzed, and provided to users on a custom basis. This took a long time to produce at a high cost, for a single purpose. By calibrating the entire data stream to the same standard in advance and by making the data accessible in a High Performance Data (HPD) structure co-located with a High Performance Computing (HPC) facility, DEA provides an enabling infrastructure for data-intensive science. DEA then organizes this calibrated data into stacks of consistent, time-stamped geographic 'tiles' so that they can be rapidly manipulated in an HPC environment. A database is used to track the data in DEA. Although DEA contains some 23 trillion individual observations, the database can be used to track every observation back to the point of collection. DEA will continually synthesize satellite images collected over the last 30 years (taken every two weeks at 25 metre squared resolution) and future images that will be taken every 5 days at 10 metre squared resolution. It will provide these images freely in a platform that can be accessed by any user, and will deliver a unique capability to process, interrogate and present this data in response to specific issues, for example water quality, land use, and forest cover in Australia.

3. USER COMMUNITY

Almost every sector in the Australian economy benefits from the use of spatial information and location technologies. Spatial information from Earth observations from space (EOS) contributes around \$5.3 billion annually through various industry programmes, and is projected to generate over 15,000 jobs by 2025. Globally, the forecasted growth of 30% per annum in geoservices provides a great opportunity for Australian companies to increase their businesses on an international scale.

Enabling the Australian spatial industry to exploit the full value of EOS information to enhance their business and be competitive in global markets is a key goal of Digital Earth Australia (DEA). The products created by Australian businesses and researchers using DEA will be transferrable to international markets as they evolve. The underpinning satellite data is global, and the United Kingdom, United

States, Canada, and South Africa are exploring their own deployments, based on DEA.

Understanding the requirements of Australian businesses for Earth observations, data infrastructure, and information products is integral to the success of DEA and to fully realizing the benefits of spatial information. In 2017/18, the DEA program will be working with the Cooperative Research Centre for Spatial Information to develop an Industry Strategy that ensures the DEA will generate value for the spatial industry and the wider Australian economy.

4. DEA PRODUCTS

The following products are initial examples of how DEA [2] will underpin innovation and capability across government, industry, and the research community.

4.1. Water Observations from Space

Water Observations from Space (WofS) uses imagery that has first been corrected for atmospheric affects, sun and sensor angles, and terrain affects. Each Landsat image is analyzed using a standard, automated algorithm to ensure each scene is analyzed in the same way. The analysis determines where water is or is not present on each image. Then, for each location, the number of water detections through time is counted and compared to the number of clear observations of that location. The final WofS product shows how often water was observed for every point in Australia in a 25 metre by 25 metre grid. Figure 1 shows the WofS filtered summary product for Australia, derived from water observations 1987 to 2014 [3].

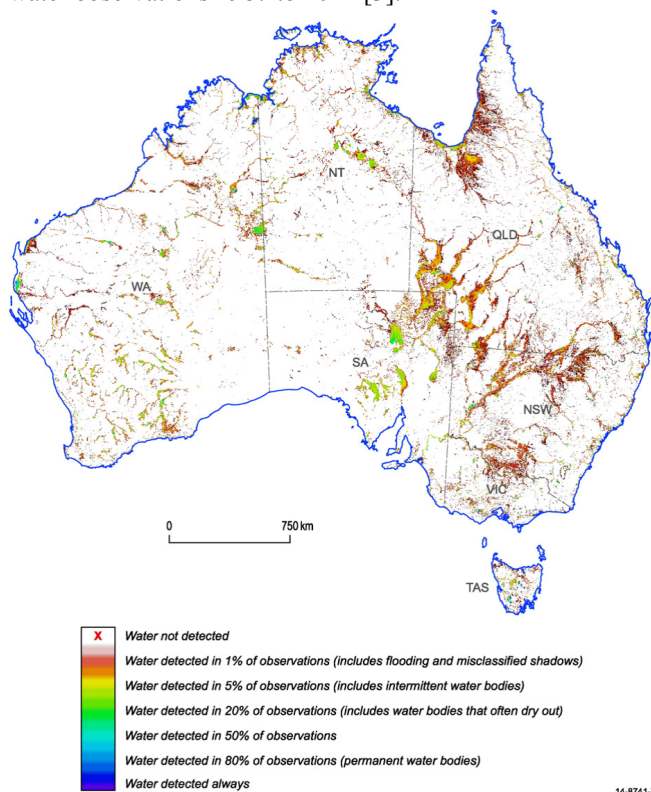


Fig. 1. The WofS filtered summary product for Australia, derived from water observations from 1987 to 2014 [3].

4.2. Fractional Cover

Fractional Cover (FC) is a measurement that splits the landscape into three parts, or fractions; green (leaves, grass, and growing crops), brown (branches, dry grass or hay, and dead leaf litter), and bare ground (soil or rock). The Fractional Cover algorithm was developed by the Joint Remote Sensing Research Program; a collaborative program that combines research, research training expertise and infrastructure from the University of Queensland's Remote Sensing Research Centre with remote sensing groups supporting the Queensland, New South Wales and Victorian governments. The detailed methodology is described in Scarth et al. (2010) [4].

DEA uses Fractional Cover to characterize every 25 m square of Australia for any point in time from 1987 to today. It provides a 25m scale fractional cover representation of the proportions of green or photosynthetic vegetation, non-photosynthetic vegetation, and bare surface cover across the Australian continent.

This information can inform a broad range of natural resource management issues; e.g. to identify large scale patterns and trends and to inform evidence based decision making and policy on topics including wind and water erosion risk, soil carbon dynamics, land management practices and rangeland condition. This information could enable policy agencies, natural and agricultural land resource managers, and scientists to monitor land conditions over large areas over long time frames.

4.3. Normalised Difference Vegetation Index

Normalised Difference Vegetation Index (NDVI) provides the ability to assess the extent of living green vegetation across the entire Australian continent at any point in time from 1987 to today. NDVI changes can also be tracked through time. Sudden drops in NDVI can be caused by a range of processes including tree clearing, cropping, or severe bushfires. Rises in NDVI can be the result of vegetation responding to increased water availability, such as crop growth or greening of irrigated pasture. NDVI changes over time can also be used to help to map different types of land cover. More gradual, multi-year trends in NDVI values can be used to identify areas where long term increases (e.g. woody weed infestation) or decreases (e.g. drought stress) in living vegetation are occurring.

4.4. Intertidal Extents Model

The Intertidal Extents Model (ITEM) is a unique new map of Australia's vast intertidal zone, the area between the land and sea that can be observed between the highest and lowest tide [5]. ITEM draws on almost 30 years of Earth observation data to map the extent and elevation profile of the intertidal zone, enabling a more realistic representation and a deeper understanding of Australia's vast coastline. The knowledge provided by ITEM can contribute to a broad range of applications, including environmental monitoring applications for migratory bird species, habitat mapping in coastal regions, hydrodynamic modelling, and geomorphological studies of features in the intertidal zone. ITEM is improving digital elevation models of Northern Australia and Queensland. Combining ITEM with other depth and elevation provides a seamless model from the deep oceans through the coastal zone to the land. Figure 2 presents the Intertidal Extents Model processing sequence. The process shown is completed for each tidal interval ensemble of tile observations. All tiles in the ensemble (regardless of data completeness or quality) are used to generate a median composite of Normalised Difference Water Index (NDWI) and a standard deviation Confidence layer. These interval based outputs are then combined to create the final products. The full details on an automated methodology to model the intertidal extent and topography of the Australian coastline, using a full time series of Landsat observations from 1987 to 2015 is described by Sagar et al., 2017 [5].

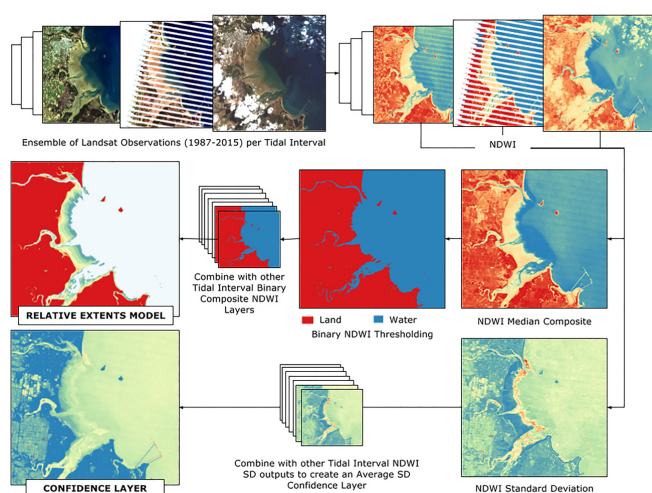


Fig. 2. The Intertidal Extents Model processing sequence [5].

4.5. Surface Reflectance

The Surface Reflectance (SR) product is the fundamental starting point for many analyses and provides the underlying data for all other DEA products at this time. The Surface Reflectance product turns the images recorded by a satellite into millions of measurements of the Earth's surface. SR is produced through a series of corrections that account for complex variations in the atmosphere, sun position, and view angle at the time each satellite image is captured, and corrects the image accordingly [6]. SR allows for a more accurate comparison of imagery captured at different times, by different sensors, in different seasons, and in different locations. It also indicates where the image has been affected by cloud or cloud shadow, contains missing data, or has been affected in other ways. These corrections have been applied to all satellite imagery from 1987 to today providing a rich history of Australia's changing landscape and coasts.

5. THE FUTURE OF DEA

Like most new and innovative technologies DEA continues to develop at the same time it is in use. Future work will include:

- Making data from more Earth observation satellites available through the DEA;
- Building new products and tools to support Australian Government agencies to better monitor, protect and enhance Australia's natural resources;
- Developing standard 'services' to support Australia's spatial industry to develop new applications, and;
- Ongoing contribution of open source code and application development to the Open Data Cube community.

6. REFERENCES

- [1] Lewis, A., et al., The Australian Geoscience Data Cube. 2017. Remote Sensing of Environment. Article in Press.
- [2] Geoscience Australia, 2017: <http://www.ga.gov.au/about/projects/geographic/digital-earth-australia>
- [3] N. Mueller, A. Lewis, D. Roberts, S. Ring, R. Melrose, J. Sixsmith, L. Lymburner, A. McIntyre, P. Tan, S. Curnow, A. Ip Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia, Remote Sensing of Environment 174, 341-352, ISSN 0034-4257.
- [4] Scarth, P., Röder, A., Schmidt, M., 2010. Tracking grazing pressure and climate interaction - the role of Landsat fractional cover in time series analysis. In: Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference (ARSPC), 13-17 September, Alice Springs, Australia. Alice Springs, NT.
- [5] Sagar, S., Roberts, D., Bala, B., Lymburner, L., 2017. Extracting the intertidal extent and topography of the Australian coastline from a 28 year time series of Landsat observations. Remote Sensing of Environment 195, 153-169.

[6] Fuqin Li, David L.B. Jupp, Medhavy Thankappan, Leo Lymburner, Norman Mueller, Adam Lewis, Alex Held, 2012. A physics-based atmospheric and BRDF correction for Landsat data over mountainous terrain. *Remote Sensing of Environment* 124, 756–770.

SENTINEL-1 DATA CUBE EXPLOITATION: TOOLS, PRODUCTS, SERVICES AND QUALITY CONTROL

I. Ali, V. Naeimi, S. Cao, S. Elefante, T. S. Le, B. Bauer-Marschallinger, W. Wagner

Department of Geodesy and Geoinformation, Vienna University of Technology
Microwave Remote Sensing Research Group, Gusshausstrasse 27–29 1040 Vienna, Austria.

ABSTRACT

This paper illustrates TU Wien's data cube approach for exploiting the Big Data volumes provided by the Sentinel-1 Synthetic Aperture Radar (SAR) mission. The Sentinel-1 data cube is established by geocoding the SAR backscatter imagery to the Equi7 grid, which is a global grid system optimized for quickly accessing and analyzing high-resolution time series. In this paper, we give an overview of TU Wien's SAR data processing chain, quality control components, data products, and discuss as one concrete application a near-real-time flood monitoring service.

Index Terms— Sentinel-1, big data, data cube, EODC, flood mapping, surface soil moisture

1. INTRODUCTION

In recent years, the volume of data acquired by earth observation (EO) satellites has been growing exponentially. Apart from the increasing number of satellites orbiting the Earth, the contribution of improved spatiotemporal resolution to the data volume is also significant. Therefore, in this age of big EO data, the domain of data driven applications is also quickly expanding [1]. Key to any operational application are scalable and highly efficient data management and processing systems.

Here we present TU Wiens data cube approach for managing and processing Sentinel-1 SAR backscatter time series on the Earth Observation Data Center (EODC) platform. The EODC platform offers an environment for EO data processing and includes a Petabyte-scale EO data archive, a high performance computing infrastructure and virtual machines for remote data handling and visualization [2].

2. SPACE-BORNE DATA VOLUME, STORAGE AND MANAGEMENT

At the EODC platform, the Sentinel-1 data are stored on both discs (for fast access) and tapes (for long-term storage and backup). Currently, 2 PB of disc space and 4 PB of tapes are available. Figure 1 illustrates the fast increase of the Sentinel-1 data volume received by and stored at the EODC. In

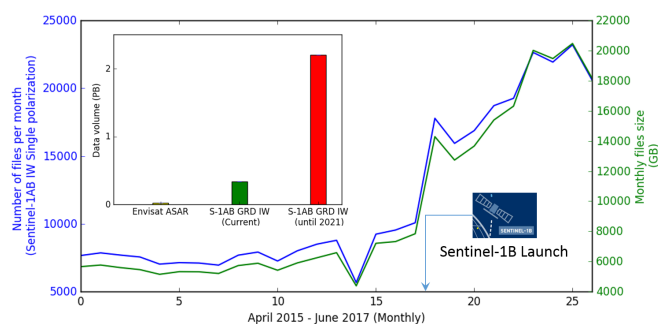


Fig. 1. Number of files and data volume of Sentinel-1 data per month. Only the statistics of the Level-1 Ground Range Detected (GRD) data products acquired in Interferometric Wide Swath (IW) mode are shown.

order to manage the raw and processed Sentinel-1 data files a dedicated meta-database (EOMDB: Earth Observation Meta-database) has been established, which allows tracking of data availability and processing status. EOMDB is being updated on regular basis so that the newly acquired scenes are become available for processing.

3. SENTINEL-1 DATA CUBE

3.1. EQUI7 grid

Recognizing the necessity of an optimized regular grid for efficient handling of large remote sensing multi-temporal data cubes, TU Wien proposed the Equi7Grid spatial reference definition for processing of global high-resolution satellite data [3]. Equi7Grid consists of seven projected continental regular sub-grids based upon an Equidistant Azimuthal projections with continent-centred projection centres, as illustrated in Figure 2.

Figure 3 shows an example of efficient spatial data handling and way how the multi-temporal data cubes physically stored using Equi7 grid. A QGIS plug-in has been developed to extract the time series of a selected pixel in a tile. Figure 3 shows an extracted Sentinel-1 backscatter time series (November, 2014 – September, 2016) of an Equi7Grid tile (E037N010T1). It is evident that handling and analysis of big

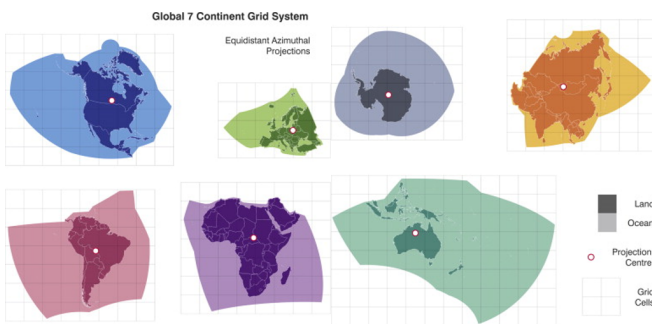


Fig. 2. Seven projected continents in Equi7 grid.

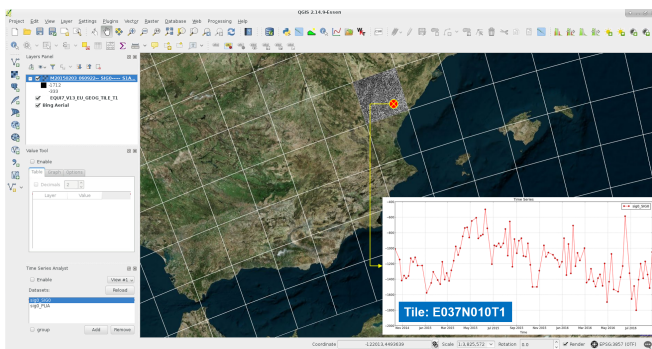


Fig. 3. An example of Sentinel-1 backscatter time series tile (E037N010T1) in Equi7 grid of a selected pixel (red point).

spatio-temporal data in cubic form is much more efficient and well organised in terms of huge data production and service development.

3.2. A strategy for efficient large scale big processing

Both EOMDB (Earth Observation Meta-database) and the Equi7Grid play a crucial roles in order to minimize the time required for management and tracking the status of big processing jobs. After the pre-processing step, where each scene is splitted into image subsets covering Equi7Grid tiles, provides an ideal scenario for further processing on a super-computer, where each computing node processes one tile. EOMDB can be queried for checking the scenes (query results in a file list) available for a give area of interest using the bounding box coordinates or shape file. The files list produced by the EOMDB can be directly used and input to SGRT workflow for pre-processing.

4. TU WIEN SAR DATA PROCESSING TOOLBOX

The SAR Geophysical Retrieval Toolbox (SGRT) is a Python-based software package developed and maintained by TU Wien for extracting geophysical parameters from Synthetic Aperture Radars (SARs) data [2]. Version 2.3.0 of the SGRT is an adaptation to Sentinel-1 of the earlier SGRT 1.0 de-

veloped for ENVISAT Advanced Synthetic Aperture Radar (ASAR) data, incorporating optimizations intended for handling the considerably higher spatial resolution and resulting explosion in data volumes foreseen of Sentinel-1 relative to ENVISAT ASAR.

Current version (v2.3.0) of SGRT consists of four components (where several workflows are defined under each component), namely: preprocessing (for pre-processing SGRT is using SNAP¹ latest beta version), analytics, production and near real-time monitoring component. SGRT is shared with EODC² cooperation partners willing to contribute to the further development of the software.

5. SAR DATA PRODUCTION

Below are the key components which are involved in SAR data production.

- *Quality control*: in the preprocessing step quality checking is a major challenge in order to remove the artefacts from image border noise. For the quality control of preprocessed data a border noise removal mask is produced [4]. For the removal of topographic artefacts in water/flood maps a global Hand (Height Above the Nearest Drainage) Index [5] mask is produced and integrated into SGRT. Finally, a dedicated workflow is developed to detect shifts after the geo-coding step in pre-processing.
- *Sample SAR products*: Figure 4 shows an example of products generated from Sentinel-1 satellite data. Water frequency (Figure 4A) is calculated from individual water maps (Figure 4B) over a given time span. Figure 4C shows an example of Sentinel-1 surface soil moisture product. Figure 4D shows the multi-temporal seasonal false colour composites using different polarizations.
- *Product validation protocol*: currently, different validation modules are being incorporated in SGRT. For example, a validation scheme for surface soil moisture product is under development.
- *Performance evaluation*: the processing of high resolution SAR data at a regional or global scale is a challenging task in terms of data storage, handing and processing. In order to process Sentinel-1 data on the order of hundreds of terabytes (TB), a high performance processing facility is inevitable [6]. Table 1 shows and overview of performance evaluation metric for preprocessing and various biophysical parameters using the Vienna Scientific Cluster (VSC-3³).

¹<http://step.esa.int/main/toolboxes/snap/>

²To join the EODC platform please contact: <https://www.eodc.eu>

³<http://vsc.ac.at>

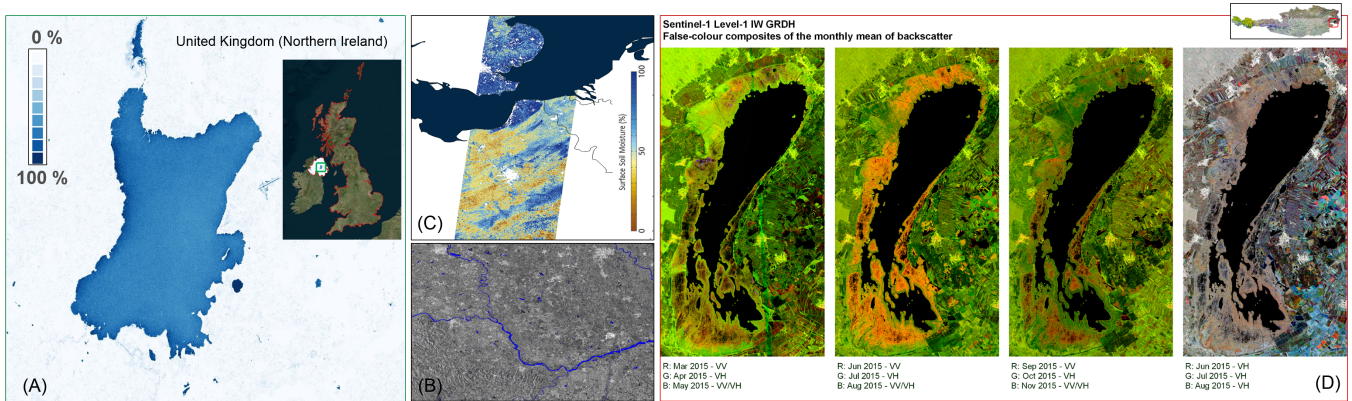


Fig. 4. (A) Flood frequency map, (B) single scene water map, (C) 1km Soil Moisture in percentage (%) and (D) multi-temporal false colour (seasonal) composites.

	Global	Europe
Monthly data volume (10m grid)	≈ 15.5 TB	≈ 4 TB
Preprocessing time (≈2 second/MB) for monthly data volume on a single node	≈ 377.3 days	≈ 96.5 days
Monthly preprocessed data volume (max: 2.5 × raw data)	≈ 39 TB	≈ 10 TB
Automatic quality check per month	≈ 3.8 days	≈ 1 day
Parameter retrieval per month	≈ 57.5 days	≈ 14.7 days
Soil moisture retrieval (500m) per month	≈ 39.1 days	≈ 4.2 days
Total processing time per month	≈ 479 days	≈ 118 days
Total storage size required per year	$(15.5 + 39) \times 12 \approx 654$ TB	$(4 + 10) \times 12 \approx 168$ TB
Total number of required node-hours per day	16 nodes	4 nodes

Table 1. Sentinel-1A&B data storage and computational resource requirement.

6. NEAR-REAL-TIME FLOOD MONITORING / MAPPING SERVICE DESIGN

In addition to operational services for surface soil moisture monitoring, we have recently implemented a "on request" flood mapping service under the framework of the I-REACT (Increasing Resilience to Emergencies through Advanced Cyber Technologies) project. The service is triggered with the request message generated from the I-REACT front-end application. The flood monitoring processor (FMP) receive the message via Azure Service Bus (messaging service between applications and services).

After receiving the message the FMP connects to the meta-database (EOMDB) and searches for the available satellite data file(s) over the target location. After getting the file list from the meta-database the flood mapping processing chain is activated and the final flood maps are pushed back to the I-REACT database called IDI (I-REACT Data Interface). Figure 5 shows an over of the service logic implemented.

7. CONCLUSIONS

We have entered in the age of big remote sensing data, and its time to fully benefit from this data. In this paper we have showed the big spatial data handling/management, processing and product generation capabilities we have developed at TU Wien and EODC, which can be further exploited to build operational service for global scale ecosystem analysis and monitoring activities. Thanks to the rapid advances in space technologies, we can now construct multi-dimensional data cubes that provide an opportunity to investigate the past assumptions that were made on basis of coarse spatial resolution and large temporal baselines. With this inflow of big-data the development of state of the art near-real time operational services are highly feasible and more accurate.

Acknowledgement

This work was supported by the Austrian Research Promotion Agency (FFG) through EOP-Danube project, "Towards

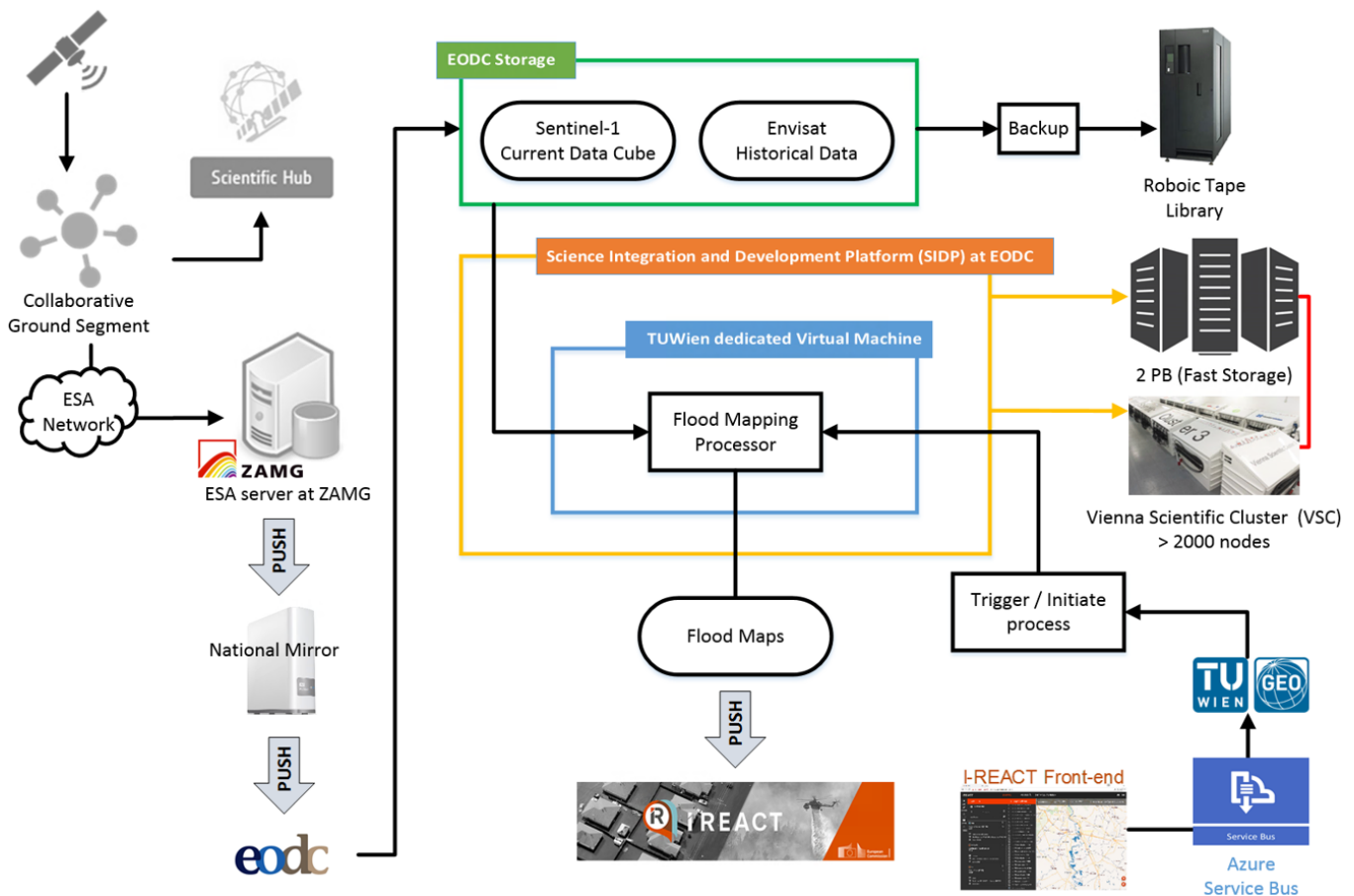


Fig. 5. An overview of service logic implementation for near-real time flood mapping.

an Earth Observation Platform for the Greater Danube Region” [854030] and the European Unions Horizon 2020 research program through I-REACT project, “Improving Resilience to Emergencies through Advanced Cyber Technologies” [700256]. The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

8. REFERENCES

- [1] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, “Big data for remote sensing: Challenges and opportunities,” *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov 2016.
- [2] V. Naeimi, S. Elefante, S. Cao, W. Wagner, A. Dostalova, and B. Bauer-Marschallinger, “Geophysical parameters retrieval from sentinel-1 sar data: A case study for high performance computing at eodc,” in *Proceedings of the 24th High Performance Computing Symposium*, San Diego, CA, USA, 2016, HPC ’16, pp. 10:1–10:8, Society for Computer Simulation International.
- [3] B. Bauer-Marschallinger, D. Sabel, and W. Wagner, “Optimisation of global grids for high-resolution remote sensing data,” *Comput. Geosci.*, vol. 72, no. C, pp. 84–93, Nov. 2014.
- [4] I. Ali, S. Cao, V. Naeimi, C. Paulik, and W. Wagner, “Methods to remove sentinel-1 border noise: Implications and importance for time series analysis,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2017 (In review).
- [5] A.D. Nobre, L.A. Cuartas, M. Hodnett, C.D. Renn, G. Rodrigues, A. Silveira, M. Waterloo, and S. Saleska, “Height above the nearest drainage a hydrologically relevant new terrain model,” *Journal of Hydrology*, vol. 404, no. 1, pp. 13 – 29, 2011.
- [6] C. A. Lee, S. D. Gasster, A. Plaza, C. I. Chang, and B. Huang, “Recent developments in high performance computing for remote sensing: A review,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 508–527, Sept 2011.

GEOSPATIAL WEB SERVICES FOR BIG CLIMATE DATA: ON-DEMAND ACCESS TO AND PROCESSING OF ECMWF'S REANALYSIS DATA

Julia Wagemann, Stephan Siemen, Sylvie Lamy-Thepaut

European Centre for Medium-Range Weather Forecasts (ECMWF), Shinfield Park, Reading RG2 9AX, UK

julia.wagemann@ecmwf.int, stephan.siemens@ecmwf.int, sylvie.lamy.thepaut@ecmwf.int

ABSTRACT

Big Earth Data brings challenges to data organisations and data users alike. Data organisations are in the need to explore new ways of storing, managing and disseminating data. Data users still struggle to exploit the full potential of Big Earth Data due to insufficient storage capacity and processing power. Geospatial web service technologies offer a time- and cost-effective way to access multi-dimensional data in a user-tailored format via the Internet and to process the data at server level. Data transport is minimised and enhanced processing capabilities are offered. In the framework of the Horizon-2020 project EarthServer-2, ECMWF set up an Open Geospatial Consortium (OGC) Web Coverage Service (WCS) for climate reanalysis data. Geospatial web services show great potential to make complex meteorological data available for non-meteorological user communities. However, geospatial web services have to fulfill specific requirements when operated operationally. In the future, large data centres have to become more progressive towards the adoption of geo-data standards and data users have to be trained using web services in order to benefit from them most.

Keywords: OGC, Geospatial web services, interoperability, climate reanalysis data, ECMWF

1. INTRODUCTION

The production of the fifth generation of ECMWF reanalysis of global climate, ERA5, has started in spring 2016. ERA5 will be the first reanalysis produced as an operational service and will be one key component of the Copernicus Climate Change Service (C3S). It will eventually have a data volume of 5 PB [3]. Our current capacity to acquire and produce geographic information with a higher temporal and spatial resolution is greater than the capacity to manage, process and analyse the data [1][2].

Large data organisations are challenged to find new solutions to effectively manage, store and archive these

massive amounts of data. They are shifting away from their pure function of storing and disseminating data. They are interested to go beyond the mere provision of data and to exploit strategies to offer services for server-based data access and processing in order to reduce the download of terra bytes of data [4]. Data users often have insufficient storage space and processing power and it became impractical to move the sheer amount of data to the researcher's local workstation. A natural consequence is the combination of data access, analytics and computing in one service [8]. However, addressing the many different user needs is a significant challenge [4].

Geospatial web services have been developed to facilitate the exchange of heterogeneous geospatial information and reveal new opportunities to disseminate, access and process data [9]. Large volumes of multi-dimensional geospatial data can be accessed via the Internet and large parts of the processing can be executed at server-side. ECMWF is part of the Horizon-2020 funded project EarthServer-2 and explores the possibilities to offer on-demand access to and server-based processing of meteorological data based on the interface standard Web Coverage Service (WCS) 2.0 [6] and its extension Web Coverage Processing Service (WCPS) [7].

We first present a list of requirements geospatial web services in general and WCS implementations in specific have to meet in order to be beneficial for both, data users and large data organisations.

We follow with a practical example from the climate science community what benefits a WCS serving climate reanalysis data may bring to data users, especially to users outside the meteorological domain.

2. REQUIREMENTS FOR GEOSPATIAL WEB SERVICES

Data users and data providers make different demands on a geospatial web service implementation and the Web Coverage Service standard specification. A full list of the collected requirements can be found in [9].

2.1. Data user requirements

Two types of data users can be distinguished: (i) the technical data user, who has technical expertise in data handling and management and (ii) the end-user, who consumes information and needs pre-processed and value-added data to make decisions. Data user requirements target mainly the first user group, who most likely have an interest to use geospatial web services in a programmatic way.

Data user requirements can be grouped in type of data request, data format, type of processing and metadata information.

- *Type of data request*: the challenge is to retrieve time-series information as well as geographical subsets in an efficient time [5]. Users of ECMWF data are specifically interested in the point (time-series) retrieval functionality.
- *Data encoding*: the WCS 2.0 standard offers image output formats, such as PNG, GeoTiff or JPEG and data output formats, such as NetCDF, JSON and CSV. Especially the latter are of interest for users, as large volumes of Big Earth Data usually need to be post-processed and users need access to the real data values, not only images.
- *Type of processing*: different Earth Science disciplines require processing operations of different kind. The climate science community, for example, is interested in generating anomalies and averages of data over a certain period of time.
- *Metadata*: besides the actual data array that is returned from a WCS request, a user would further expect accompanied axes information for each data value, as this facilitates further data processing and visualization.

2.2. Data provider requirements

Data provider's requirements rely on both the WCS standard interface and the WCS server implementation. ECMWF as an operational data center that disseminates data in a 24/7 mode brings in specific requirements. Requirements can be grouped into data formats, coordinate systems, data semantics, service performance, scalability and Quality of Service.

- *Data formats*: the support of native data formats, especially NetCDF and GRIB, as input data formats is critical for data providers. This allows data users not to deal with community-specific data formats and are able to directly access the data values. Especially the proprietary data format for meteorological data, GRIB, is challenges for users outside the meteorological community.
- *Coordinate systems*: support of different Coordinate Reference Systems (CRSs), beyond the standard WCS84 CRS is important. NWP models e.g. often use a reduced Gaussian grid, which is regular in longitude

and reduces the number of grid points along the shorter latitude lines near the poles.

- *Data semantics*: a data provider should be able to manually set up the data model, including multiple numbers of axes and the specification if axes have a continuous or discrete space.
- *Performance (time of response)*: the response time for WCS requests should be minimal. The challenge here is to have a similar response time for different types of data requests, be it geographical sub-setting or point data retrieval [5].
- *Scalability*: scalability in the context of web services is considered as being equally performant for 100,000 users as for ten users per day.
- *Quality of Service (QoS)*: offering a Web Coverage Service in an operational way, all Quality of Service aspects, such as performance, scalability, reliability and availability, play an important role. A dedicated working group at OGC addresses this very important issue to define standard QoS metrics.

3. BENEFITS OF A WCS FOR THE CLIMATE SCIENCE COMMUNITY

ECMWF ERA-interim data (and soon its successor ERA5) are currently the best representation of the historic state of the atmosphere. One ERA-interim parameter (temporal/spatial resolution: 6-hourly, 0.5 deg lat/lon grid) from 1 January 1979 to 31 December 2014 has a volume of 27 GB.

There are currently two options to retrieve ERA-interim data from ECMWF's Meteorological Archival and Retrieval System (MARS):

- manually, via ECMWF's Web User Interface (<http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/>), or
- programmatically, via ECMWF's Web-API (<https://software.ecmwf.int/wiki/display/WEBAPI/ECMWF+Web+API+Home>)

Both options allow the efficient download of data in either GRIB or NetCDF format. A user can further retrieve either a global field or a geographical subset thereof.

Data access via a WCS and processing via a WCPS provide additional flexibility to the common data retrieval workflow:

- 1) With the help of a WCS request, additionally to geographical subsets, time-series information for one specific latitude/longitude information can efficiently be retrieved without requiring a bulk download of multiple 2D fields.

- 2) Mathematical condenser functions provided by a WCPS offer additional processing capabilities and are able to condense a 3D raster stack into its 2D average, minimum or maximum. Through the SQL-like query language, mathematical operations can be applied. Data values of 2m air temperature, for example, can be converted from Kelvin to degree Celsius on-the-fly.

```

http://earthserver.ecmwf.int/rasdaman/ows?
Service=WCS&version=2.0.1
&request=ProcessCoverages
&query=
for c in (temp2m) return encode
(c[Lat(41.9), Long(12.5), ansi("2002-01-
01T00:00":"2002-01-31T18:00")]-273.15,
"csv")
    
```

Example of a WCPS request for 2 m air temperature of Rome for January 2002, converted from Kelvin to degrees Celsius and returned in csv format.

- 3) A WCS for a meteorological data offers independency of the data format from a data user perspective. Data users do not have to deal with complex and community-specific data formats, such as GRIB, and have direct access to the real data. This increases the data uptake from users outside the meteorological community to a large extent.

4. REDEFINING THE GEOSPATIAL DATA ANALYSIS WORKFLOW

Standardized geospatial web services, such as Web Coverage Service (WCS), are an effective way to redefine the geospatial data analysis workflow. A WCS is the link between a large-scale data provider and a technical data user (Figure 1). The data provider’s responsibility, such as ECMWF, is to offer, maintain and manage data via a WCS. Technical data users can then analyse data directly without prior download. They can integrate data access into their scientific processing routines or can build web applications with the help of web frameworks and Javascript libraries. Technical data users constitute the link between data provider and the end user, who can be decision-makers or government agencies. Users, who need value-added information retrieved from GBs to TBs of data.

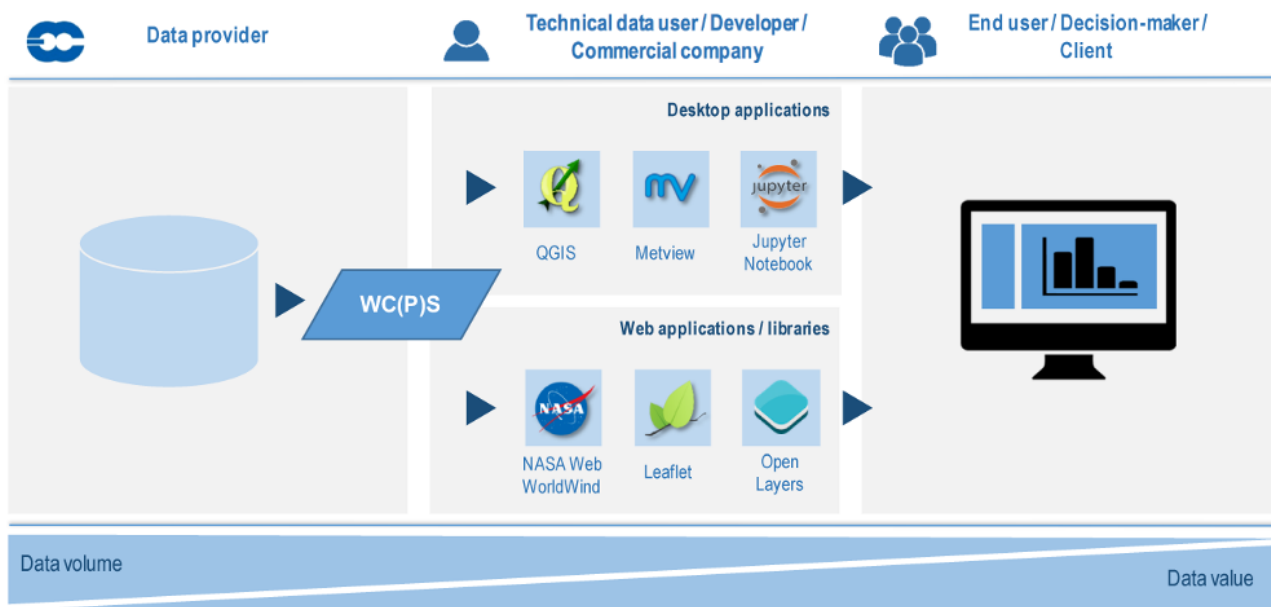


Figure 1) A Web Coverage Service as part of the Geospatial data analysis workflow

5. CONCLUSION AND OUTLOOK

Geospatial web services, especially the interface standard Web Coverage Service 2.0, bring new opportunities to access large volumes of meteorological and climate data via the internet and to process them at server-side. Users benefit, as they are not required to download massive amounts of data anymore. WCS requests can directly be integrated into existing processing routines, giving users more time to analyse and interpret the data. In order to truly benefit from geospatial web services and interoperable data standards, large data centers have to become more progressive in their implementation. At the same time, domain-specific requirements for these standard protocols have to be defined. An example for WCS is the Earth Observation Application profile, that defines additional requests specifically for EO data [6]. A similar application profile is under development for the MetOcean domain.

While new data services are being developed, data users have to be trained in using them. Training has to show the benefits for users in order to build up trust in these services.

The overall goal are service federations that combine access and processing of data from different WCSs into one single request. This would lead to a true interoperability of decentralized data repositories worldwide.

REFERENCES

- [1] A. Dasgupta, “The Continuum: Big Data, Cloud & Internet of Things”, *Geospatial World* 7(1):14-20, 2016
- [2] H.L. Guo, L. Wang and D. Liang, “Big Earth Data from Space: A New Engine for Earth Science”, *Science Bulletin* 61(7): 505-513, 2016
- [3] H. Hersbach, D. Dee, “ERA5 Reanalysis is in Production”, *ECMWF Newsletter* 147: 7, 2016
- [4] E.J. Keans, T.R. Karl, M.D. Tanner, J.J. Bates, J.L. Privette, W.J. Glance, X. Zhao, “On the preservation and application of climate data records in a big data world”, *Proceedings of BiDS14* (Conference on Big Data from Space 2014), ESA-ESRIN, Frascati Italy, pp. 1-3, November 2014
- [5] K. Kyzirakos, S. Manegold, M. Kersten, “Scientific databases, the new substrate for Earth Observation”, *Proceedings of BiDS16* (Conference on Big Data from Space 2016), Santa Cruz, Tenerife, pp. 27-30, March 2016
- [6] OGC (Open Geospatial Consortium), “OGC WCS 2.0 Interface Standard: OGC Document 09-110r4.”, 2012, <http://www.opengeospatial.org/standards/wcs> (last access: 15 October 2017)
- [7] OGC (Open Geospatial Consortium), “Web Coverage Processing Service (WCPS) Language Interface Standard: OGC Document 08-068r2.”, 2009, <http://www.opengeospatial.org/standards/wcps> (last access: 15 October 2017)
- [8] G.L. Potter, T. Lee, L. Carriere, “Improving access to climate model, observational, and reanalysis data”, *Proceedings of BiDS14* (Conference on Big Data from Space 2014), ESA-ESRIN, Frascati Italy, pp. 86-89, November 2014
- [9] J. Wagemann, O. Clements, R.M. Figuera, A.P. Rossi and S. Mantovani, “Geospatial web services pave new ways for server-based on-demand access and processing of Big Earth Data”, *International Journal of Digital Earth*: 1-19 (published online), 17 Jul 2017

ACKNOWLEDGEMENTS

This work was supported by the European Union’s Horizon 2020 Framework Programme research and innovation agreement [grant number 654367]

THE E-SENSING ARCHITECTURE FOR BIG EARTH OBSERVATION DATA ANALYSIS

Gilberto Camara, Gilberto Queiroz, Lúbia Vinhas, Karine Ferreira, Ricardo Cartaxo, Rolf Simoes, Eduardo Llapa, Luiz Assis, Alber Sanchez

National Institute for Space Research (INPE), Earth Observation Directorate
São José dos Campos, SP, Brazil

ABSTRACT

This work presents an architecture for big Earth Observation data analytics. It uses array databases to support storage and management of large volumes of satellite image time series. The analysis methods are developed in R and enable using the full depth of satellite image time series with advanced statistical learning algorithms. New kinds of web services allow data access and remote data processing of time series. The *e-sensing* architecture has been designed with a focus on land use and land cover classification using SITS, an area of Earth observation where much progress is required. This architecture is fully implemented and has already allowed innovative results in land use and land cover mapping. The method works with big data sets with a minimal set of assumptions to increase its generality. Our work promotes reproducibility and reuse of the methods and results.

Index Terms— Earth observation, web services, satellite image time series, array databases, science reproducibility, open source.

1. INTRODUCTION

The data deluge resulting from the open access policies for Earth observation (EO) data has brought about a major challenge: *How to design and build technologies that allow the EO community to analyse big data sets?*. Developing such a solution is hard because current technologies for big data management are quite different and incompatible. Alternatives include using flat files [1], MapReduce-based solutions such as Google Earth Engine [2], and distributed multidimensional array databases such as Rasdaman [3] and SciDB [4]. Each choice has its advantages and drawbacks, and fits certain needs better than others.

The first option of an infrastructure for big EO data is to store EO data as flat files and use file management systems. This is the approach taken by the Australian Data Cube [1]. This choice makes it easy to preprocess images from different sources so that they become geometrically and radiometrically

This work is supported by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014–08398–6) and by Germany’s International Climate Initiative (IKI/BMUB) under grant 17-III-084-Global-A-RESTORE+. Gilberto Camara is also supported by CNPq (grant 312151–2014–4).

compatible. Data merging and cross-calibration tasks are simple to perform. Existing pixel-based image analysis methods can be applied to big data sets. However, these simple infrastructures have a high management cost. Data analysis proceeds by searching all the relevant files. The programs open each file, extract the relevant data and then move onto the next file. When all the relevant data has been gathered in memory, the program can begin its analysis. Working with time series becomes specially burdensome because of the number of files that must be opened for a single time series to be retrieved. Managing 10,000 - 100,000 files at once can lead to scalability and performance bottlenecks.

An alternative is to take a mainstream solution used for other big data applications and adapt it to EO data. This is the case of MapReduce-based solutions such as Google Earth Engine [2]. The MapReduce model has been motivated by highly parallel applications such as text queries and there are open source implementations such as Spark. MapReduce architectures are very efficient for problems where each pixel is processed independently. They lack flexibility for big EO analytics, since they use an excessive granularity when breaking the data into parts. Region-based methods such as image segmentation are not supported, nor large-scale time series analysis are possible.

A third option is to use array databases such as Rasdaman [3] and SciDB [4]. Array DBMS reduce the impedance mismatch between the data model (raster), the storage model (arrays) and analysis functions such as linear algebra and image processing. These databases split large volumes of data in distributed servers in a “shared nothing” way. Each server controls its local data storage. Arrays are multidimensional and uniform, as each array cell holds the same user-defined number of attributes. Array databases allow organising EO data to meet the needs of different applications. Comparative studies show the SciDB architecture to be more efficient and more flexible for processing remote sensing data than MapReduce [5]. However, since array databases are designed for scientific data management, there is much less experience with them. Developers using SciDB have to spend significant effort for system configuration and performance tuning. Despite these problems, we consider array databases to be the best choice for support innovative big EO data analytics.

One of the areas where array DBMS allow advances on big EO data analytics is when processing dense satellite image time series (SITS). Using SITS is a leading research trends in Remote Sensing [6], [7]. One of the more promising applications of SITS is measuring land use change. Land use change is important for Brazil, one of the world's largest agricultural producers with one of Earth's richest biodiversities. Many researchers have also pointed out the need for improving future global land cover products [8], [9]. Given this motivation, the *e-sensing* architecture has been designed with a focus on land use and land cover classification using SITS.

This work presents innovative methods for using the full depth of satellite image time series for extracting information from big Earth observation data. We have developed a full open source architecture that allows efficient processing of large-scale data sets, coupled with advanced data analytic methods. Our focus is on extracting the most information from dense time series of remote sensing satellites such as MODIS, LANDSAT, and SENTINEL, or combinations of those.

2. DESIGN DECISIONS

The *e-sensing* architecture has been designed with a different perspective than other proposals for Earth Observation Data Cubes [1]. We believe the gains of using big EO data will come from new analytical methods, and our design reflects such aim. A key decision for big EO architectures is the choice of programming environment. We chose R, which has more than 11,000 packages for statistical computing and graphics, including spatial analysis, time-series analysis, classification, clustering, and machine learning. Using R, it is easier for researchers to develop new methods and to collaborate with their peers. SciDB has a streaming interface that runs R scripts in parallel directly on each server (Figure 1). Combining array DBMS with R statistical computing is a natural solution for EO applications, allowing a good balance between massive parallel data processing and maximum flexibility in algorithm design.

Scientists also need tools for small-scale testing and for scaling up their work. We developed two web services to support these tasks [10]. The Web Time Series Service (WTSS) retrieves time series of Earth observation data for specific locations. The Web Time Series Processing Service (WTSPS) enables users to run R scripts on data cubes of Earth Observation data. These Web Services enable scientists to test their analysis methods first on their desktops and then move them to big EO data cubes.

Based on these considerations, the *e-sensing* architecture uses the following building blocks:

- 1) The SciDB open source array database [4] that allows easy mapping of big EO data to its data structure.

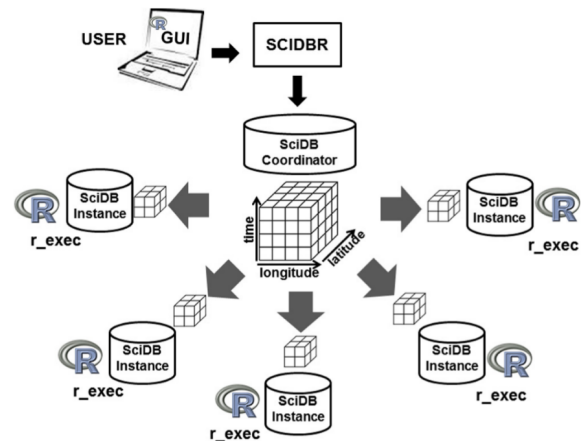


Fig. 1: Remote execution of R scripts in SciDB

- 2) R as the tool for big data analytics, so that researchers can thus scale up their methods, reuse previous work, and collaborate with their peers.
- 3) The R packages SITS [11] and dtwSat [12], for big EO analytics on satellite image time series.
- 4) Web services (WTSS and WTSPS) for big EO data, adapted to the needs of satellite image time series [10].
- 5) The architecture is fully open source, being made available online at <https://github.com/e-sensing/>.

3. MATCHING DATA INFRASTRUCTURES TO ANALYTICAL NEEDS

Most studies on time series for land cover classification in the literature use classical remote sensing methods [6]. For multiyear studies, researchers derive “best-fit” yearly composites and then classify each composite image separately. The results from different periods are compared to detect change. We denote these works as taking a *space-first, time-later* approach.

Space-first, time-later methods do not use the full potential of remote sensing time series. The benefits of SITS increase when the temporal resolution of the big data set captures the most important changes. In these cases, the temporal autocorrelation of the data will be stronger than the spatial autocorrelation. Given data with adequate repeatability, a pixel is more related to its temporal neighbours than to its spatial ones. In these cases, *time-first, space-later* methods lead to better results than the *space-first, time-later* approach.

There has been much recent interest in the Earth observation community on using advanced statistical learning methods such as support vector machines [13] and random forests [14]. However, most researchers still use a *space-first, time-later* approach in connection with these methods. The dimensions of the decision space are limited to the number of spectral bands or their transformations. These approaches do

not use the power of advanced statistical learning techniques to work on high-dimensional spaces and with big training data sets [15].

The analytical methods of the *e-sensing* architecture combine data from image time series with statistical learning, using a *time-first, space later* approach. These methods use the full depth of dense time series to train advanced predictive models. These model include linear and quadratic discrimination analysis, support vector machines, random forests and neural networks. In a typical classification problem, we use time series with known land cover labels to derive measures that capture class attributes. Based on these measures, referred as training data, we provide support to select a predictive model that allows inferring classes of a larger data set.

Our proposal uses the full depth of satellite image time series to create large dimensional spaces. The method we developed has a deceptive simplicity: *use all the data available in the time series samples*. The idea is to have as many temporal attributes as possible, increasing the dimension of the classification space. Our experiments found out that modern statistical models such as support vector machines, and random forests perform better in high-dimensional spaces than in lower dimensional ones.

To illustrate the approach, Figure 2 shows the plot of the NDVI values of 370 time series for land cover class "Pasture", based on ground samples. Each thin line is one time series. The darker lines are the median and first and third quartile values. By visualizing the data, the challenge of distinguishing noise from natural variation becomes clear. The data shows natural variability due to different climate regimes and shows noise associated to cloud cover. To avoid losing information, we use the raw data such as this one to train a support vector machine, a classifier which is robust to noisy data sets.

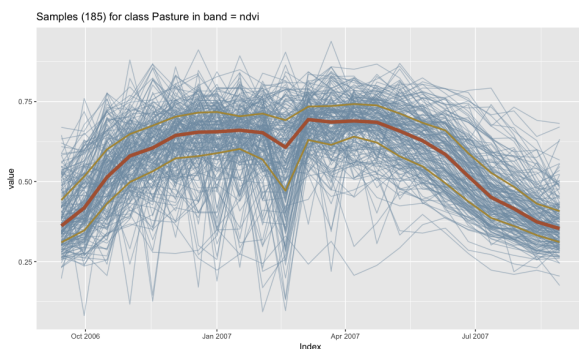


Fig. 2: Time series of 370 ground samples for land cover class "Pasture" in the state pf Mato Grosso, Brazil (source: authors).

As a case study, we developed a detailed land use change map of the state of Mato Grosso, Brazil, an area of 900,000 km², which has about 20 billion time series measures. We

used the MODIS MOD13Q1 product from 2001 to 2016, provided every 16 days at 250-meter resolution, with 23 samples per year. By taking samples of labelled time series with 4 bands, we feed the statistical inference model with a 92-dimensional attribute space. For the analysis, we used the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), and the near infrared (NIR) and middle infrared (MIR) bands. We defined nine classes (see Table 1 that include the most important crops and production systems in Mato Grosso. Based on a 5-fold cross validation, we estimate an overall accuracy of 94% and the Kappa index was 0.92. Producer's and user's accuracies of all classes were close to or better than 90%. This confirms the applicability of the proposed method in classify agricultural areas. In general, results show good discrimination between different crops, which improves on previous work [16], [17], [18].

Table 1: Confusion matrix of MODIS time series images, obtained by 5-fold cross validation of classification of field data, and values of producer's accuracy (PA) and user's accuracy (UA) for each class.

	1	2	3	4	5	6	7	8	9	UA
1 Cerrado	393	0	0	12	0	0	0	0	0	0.97
2 Fallow-Cotton	0	33	0	0	1	2	0	0	0	0.92
3 Forest	1	0	136	0	0	0	0	0	0	0.99
4 Pasture	6	0	1	357	3	1	0	5	0	0.96
5 Soy-Corn	0	1	1	1	352	18	0	26	4	0.87
6 Soy-Cotton	0	0	0	0	13	376	0	4	0	0.96
7 Soy-Fallow	0	0	0	0	0	0	88	0	0	1.00
8 Soy-Millet	0	0	0	0	25	2	0	199	2	0.87
9 Soy-Sunflower	0	0	0	0	4	0	0	1	47	0.90
PA	0.98	0.97	0.99	0.96	0.88	0.94	1.00	0.85	0.89	

4. COMPUTING PERFORMANCE

The architecture has been implemented operationally at Brazil's National Institute for Space Research. In terms of hardware, our architecture uses 2 clusters. Each cluster has 5 servers with 2 CPUs with 6-cores each, operating at 2.4GHz with a 15MB cache. Each server has 96 GB of RAM, and 16 TB of data storage. This gives 60 cores per cluster that can work in parallel in a "shared-nothing" data storage. The array database SciDB includes the full set of MODIS MOD09Q1 images at 250 meter resolution for South America, with 13,800 images associated to 317 billion data series. It also include selected datasets of mixed LANDSAT-8 and MODIS data sets, at 30 meter resolution.

In terms of performance, the classification scales up almost linearly. The full processing of all time series to classify 16 years of data in Mato Grosso state (900,000 km²) takes about 6 hours using the R-SciDB interface. We also processed all of the area of Brazil's Cerrado biome (2,050,000 km²) in about 13 hours. This shows that distributed processing with a right degree of granularity can compensate for the slower

performance of R scripts, compared with compiled languages. By using R, researchers have much flexibility when designing data analysis methods. Given these results, we argue that using SciDB combined with R is an adequate solution for big Earth Observation data analytics.

Table 2: Performance time for selected case studies

Case Study	Area (km ²)	Data dimensions	Measures (millions)	Proc time (hours)
Mato Grosso	900,000	92	20,000	6
Cerrado	2,050,000	92	50,000	13

5. FINAL REMARKS

This paper discusses the design of an architecture that allows using satellite image time series with advanced statistical learning. Its results indicate that solutions based on array DBMS, R algorithms, and dedicated web services are well suited for satellite image time series analysis. This knowledge platform expands what can be done with big EO data, allowing scalability and reproducibility, without major compromises in performance. In the long run, it shows that the *time-first, space later* approach is an important complement of more traditional image analysis methods.

Combining array databases with R statistical computing is not an universal solution for big Earth observation data analysis. Alternative designs such as the Australian Data Cube (flat files) and Google Earth Engine (MapReduce) provide support for important studies in cases where the analysis methods are established and the novelty comes from applying them to big data. In areas where the current methods are not adequate and progress is required, such as global land cover, it is important to design new architectures such as the one proposed in the paper. We hope that our results encourage further work on the use of satellite image time series for land cover classification.

6. REFERENCES

- [1] A. Lewis, S. Oliver *et al.*, “The Australian Geoscience Data Cube — Foundations and lessons learned,” *Remote Sensing of Environment (online)*, 2017.
- [2] N. Gorelick, M. Hancher *et al.*, “Google Earth Engine: Planetary-scale geospatial analysis for everyone,” *Remote Sensing of Environment*, 2017.
- [3] P. Baumann, A. Dehmel *et al.*, “The multidimensional database system RasDaMan,” *ACM SIGMOD Record*, vol. 27, no. 2, pp. 575–577, 1998.
- [4] M. Stonebraker, P. Brown *et al.*, “SciDB: A database management system for applications with complex analytics,” *Computing in Science & Engineering*, vol. 15, no. 3, pp. 54–62, 2013.
- [5] K. Doan, A. O. Oloso *et al.*, “Evaluating the impact of data placement to Spark and SciDB with an Earth Science use case,” in *2016 IEEE International Conference on Big Data*, 2016, pp. 341–346.
- [6] C. Gomez, J. C. White, and M. A. Wulder, “Optical remotely sensed time series data for land cover classification: A review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55 – 72, 2016.
- [7] V. J. Pasquarella, C. E. Holden *et al.*, “From imagery to ecology: leveraging time series of all available LANDSAT observations to map and monitor ecosystem state and dynamics,” *Remote Sensing in Ecology and Conservation*, vol. 2, no. 3, pp. 152–170, 2016.
- [8] S. Fritz, L. See *et al.*, “Highlighting continued uncertainty in global land cover maps for the user community,” *Environmental Research Letters*, vol. 6, no. 4, p. 044005, 2011.
- [9] N. Tsendbazar, S. de Bruin, and M. Herold, “Assessing global land cover reference datasets for different user communities,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, no. Sup C, pp. 93 – 114, 2015.
- [10] L. Vinhas, G. Ribeiro *et al.*, “Web services for big Earth observation data,” in *Proceedings of the 17th Brazilian Symposium on GeoInformatics*. Campos do Jordão, SP, Brazil: INPE, 2016, pp. 26–35.
- [11] R. Simoes, G. Camara *et al.*, *SITS: Satellite Image Time Series Analysis*, 2017, r package version 0.9.30. [Online]. Available: <https://github.com/e-sensing/sits/>
- [12] V. Maus, G. Camara *et al.*, “dtwSat: Time-Weighted Dynamic Time Warping for Satellite Image Time Series Analysis in R,” *Journal of Statistical Software (accepted)*, 2017.
- [13] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.
- [14] M. Belgiu and L. Dragut, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.
- [15] G. James, D. Witten *et al.*, *An Introduction to Statistical Learning: with Applications in R*. New York, EUA: Springer, 2013.
- [16] J. Kastens, J. Brown *et al.*, “Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil,” *PLOS ONE*, vol. 12, no. 4, p. e0176168, 2017.
- [17] M. N. Macedo, R. S. DeFries *et al.*, “Decoupling of deforestation and soy production in the southern Amazon during the late 2000s,” *PNAS*, vol. 109, no. 4, pp. 1341–1346, 2012.
- [18] D. Arvor, M. Jonathan *et al.*, “Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil,” *International Journal of Remote Sensing*, vol. 32, no. 22, pp. 7847–7871, 2011.

MASS PROCESSING OF SENTINEL-1 AND LANDSAT DATA FOR MAPPING HUMAN SETTLEMENTS AT GLOBAL LEVEL

Corbane C.¹, Pesaresi M.¹, Politis P.², Syrris V.³, Florczyk J. A.¹, Soille P.³, Maffenini, L.⁴, Burger A.³, Vasilev V.³, Rodriguez D.³, Sabo F.⁵, Dijkstra L.⁶ and Kemper T.¹

¹ European Commission, Joint Research Centre (JRC), Directorate for Space, Security & Migration, Italy

² Arhs Developments S.A., 2b, rue Nicolas Bové, L-1253 Luxembourg

³ European Commission, Joint Research Centre (JRC), Directorate for Competences, Italy

⁴ GFT Italia S.r.l., Via Campanini, 6, 20124 Milano, Italy

⁵ Arhs Developments Italia S.r.l., Via Privata Fratelli Gabba no. 1/A, 20121, Italy

⁶ European Commission, Directorate General for Regional and Urban Policy, Belgium

ABSTRACT

Continuous global-scale mapping of human settlements in support to international agreements calls for massive volumes of multi-source, multi-temporal and multi-scale Earth Observation (EO) data. In this paper, the latest developments in terms of processing EO datacubes for the purpose of improving the Global Human Settlement Layer (GHSL) data are presented. Two workflows with Sentinel-1 and Landsat data collections were run leveraging on the Joint Research Centre (JRC) Earth Observation and Processing Platform (JEODPP). The paper presents the processing workflows and the results of the two main workflows giving insights into the enhanced mapping capabilities gained by analyzing Sentinel-1 and Landsat datasets, and the lessons learnt in terms of handling and processing Earth Observation datacubes.

Index Terms— Big data, Global Human Settlement Layer, Sentinel-1, Landsat, JRC Earth Observation Data and Processing Platform

1. INTRODUCTION

Knowledge of the global distribution and evolution of human settlements has become one of the key requirements for monitoring progress towards sustainable development of urban and rural areas. In this regard, Earth Observation (EO) is recognized as a substantial enabler of informed decision-making by allowing measuring and monitoring the agreed objectives. Currently there are several EO-derived or EO-supported maps of built-up areas. The most exhaustive and widely used maps of built-up areas are those produced with TerraSAR-X data, namely the Global Urban Footprint (GUF) [1] or with the free Landsat data such as the GlobeLand30 (GLC-30) [2] and the Global Human Settlement Layer (GHSL) [3]. Compared to its concurrent maps of human settlements, the GHSL currently represents the most up-to-date and multi-temporal gridded dataset on the physical characteristics and the dynamics of human settlements, fully supporting the concept of open data. Drawing on 40 years of Landsat data collections, multi-temporal GHSL grids describing built-up areas been produced for the periods 1975, 1990, 2000 and 2014. In

order to keep up with the challenges related to the reporting framework of the international agreements, a continuous global-scale mapping of GHSL is required. This calls for massive volumes of EO data with the following characteristics: worldwide coverage, multi-source, multi-temporal, multi-scale, high dimensional, highly complex and unstructured. To meet the demands of global-scale mapping of human settlements from space, not only mass storage infrastructures are needed but also datacubes and novel data analytics combined with high-performance computing platforms should be designed. This paper presents the latest developments in terms of processing big data for the purpose of mapping human settlements from space. The main driver of this work is to give insights into methodological aspects of the processing of Sentinel-1 and Landsat data within the European Commission JRC Earth Observation Data and Processing Platform (JEODPP). A quantitative analysis shows how the design and use of adaptive datacubes can improve the mapping of built-up areas within the GHSL scope.

2. METHOD AND DATASETS

The GHSL production workflow builds on the Symbolic Machine learning (SML) method that was designed for remote sensing big data analytics [4]. The SML classifier automatically generates inferential rules linking the image data to available high-abstraction semantic layers used as training sets. The SML schema is based on two relatively independent steps:

(1) Reduce the data instances to a symbolic representation (unique discrete data-sequences);

(2) Evaluate the association between the unique data-sequences subdivided into two parts: X (input features) and Y (known abstraction derived from a learning set).

In the application proposed here the data-abstraction association is evaluated by a confidence measure called ENDI (evidence-based normalized differential index) which is produced in the continuous $[-1, 1]$ range [4].

Using the SML classifier two large scale workflows for automatic extraction of built-up areas from Sentinel-1 and Landsat data were deployed and executed within the JEODPP [5], [6]:

1) The purpose of the first workflow with Sentinel-1 data was to produce up-to-date information on human settlements while mitigating commission and omission rates of the first GHSL multitemporal product derived from Landsat data collections [3],[7]. A worldwide monotemporal coverage of Sentinel-1A data Ground Range Detected data was queried, downloaded and processed on the JEOPDD. The Sentinel-data has a total volume of 8.1TB, and consists of 5,026 Sentinel-1A images.

2) The second workflow was meant to evaluate the advantages of incremental learning of the SML classifier. Hence, the Landsat data collections were reprocessed and several tests were implemented during which both the artificial surfaces from GLC-30 and the built-up areas derived from Sentinel-1 (results from the first workflow) were injected as learning sets within the SML workflow. In total, 32,808 images with a total volume of 20 TB organized in four data collections centered at 1975, 1990, 2000 and 2014 were processed.

3. ADAPTIVE DATACUBES

All input datasets are combined within adaptive datacubes, in which not only the images from Sentinel-1 and Landsat, but also the semantic layers used for learning the SML are captured in a “space-time-abstraction level” series of tiles that share an identical geospatial footprint and reference grid. The concept of adaptive datacubes implemented on the JEODPP corresponds to data/information layers, which are linked by artificial intelligence or models aiming to converge to an improved representation of the reality. This is achieved through evolutionary, machine learning and classification actions on heterogeneous, partially-consistent, large mass of data recorded by different sensors in different time instances. The key features of the modular layered architecture of the adaptive datacubes are illustrated in Fig.1.

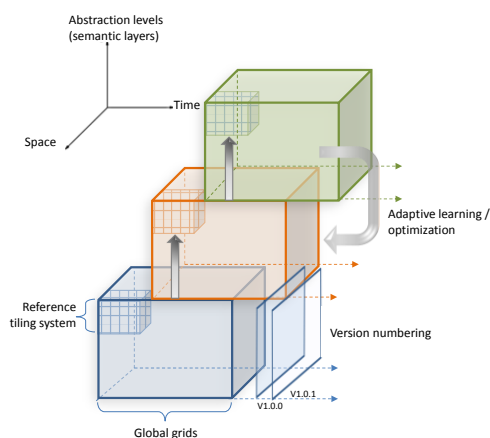


Fig.1 Adaptive datacubes implemented on the JEODPP in the context of the GHSL

Three main types of datacubes co-exist in this design:

- 1) Gridded sensor data including optical and radar data (i.e. Landsat data collections, Sentinel-1 images). The images are stacked according to the time of data capture.
- 2) Thematic data used for the analysis of earth observation data which may consist of: i) a low abstraction layer directly derived from the first family of datacubes. This pertains mainly to cloud and shadow masks or mosaics of the earth observation data (e.g. the global mosaic of Sentinel-1 ([8]); ii) ancillary data used for processing, analyzing and validating the first family of datacubes (e.g. topographic features derived a digital elevation model, global land cover data like the GLC-30).
- 3) Hierarchical semantic layers that translate the sensor data into higher level terms with degrees of abstraction: simple spectral indices like the Normalized Difference Vegetation Index (NDVI) correspond to the lowest abstraction level. While image descriptors like textural and morphological features are considered as medium abstraction semantic layers. The outputs of the SML (in terms of both encoded data sequences and ENDI confidence measures) are considered as the top level abstraction semantic layers.

All three data types of datacubes are implemented as regular tiles with a global coverage, hence establishing a hyperlattice-like structure. The global grid consists of 71, 556 tiles of 150 km × 150 km. The software environment used to manage and interact with the data while supporting high performance computing environment is provided by the JEODPP. The most distinct feature of this architecture is its ability to dynamically adapt to evolving or changing variables. As input sensor data increase in volume and as calibration parameters as well as ancillary inputs mature and improve, the reprocessing of data is required to maintain the products suite as best practice. This is effectively handled by the mechanism of adaptive learning and optimization, embedded in the SML classifier, combined with version numbering and update control schemas. Since all three families of datacubes are chained, with one datacube used as input to the next, ubiquitous updates and incremental improvements of the GHSL production are made possible thanks to this layered architecture.

4. RESULTS

A comparative analysis of the results of built-up areas extraction from the different sensors and workflows was performed. In the absence of appropriate global reference data, built-up areas obtained from the GUF product were used as a reference for cross-comparison. Three products are compared here using the GUF built-up areas as benchmark: - the first multitemporal product derived from Landsat data (GHS BU LDSMT v.2015);

- the Sentinel-1A based product (GHS BU S1 2016) product resulting from the first workflow and;
- the new product obtained from the reprocessing of Landsat data in the second workflow (GHS BU LDSMT v. 2017).

Standard accuracy and respective error metrics derived from the confusion matrix were calculated ([9], [10]). Given the lack of a single universally accepted measure of agreement, we use here a combination of three main metrics to give a complete picture of the differences among the products: the kappa coefficient, the commission and the omission errors.

Fig.2 shows the results of the comparative analysis at the global level reported by continent as a way to consider the regional variability in settlements characteristics in terms of building material construction type, structure and physical surrounding. The median and standard deviation values of the three accuracy metrics were calculated for 23,134 tiles of 150 × 150 km size covering all the landmass.

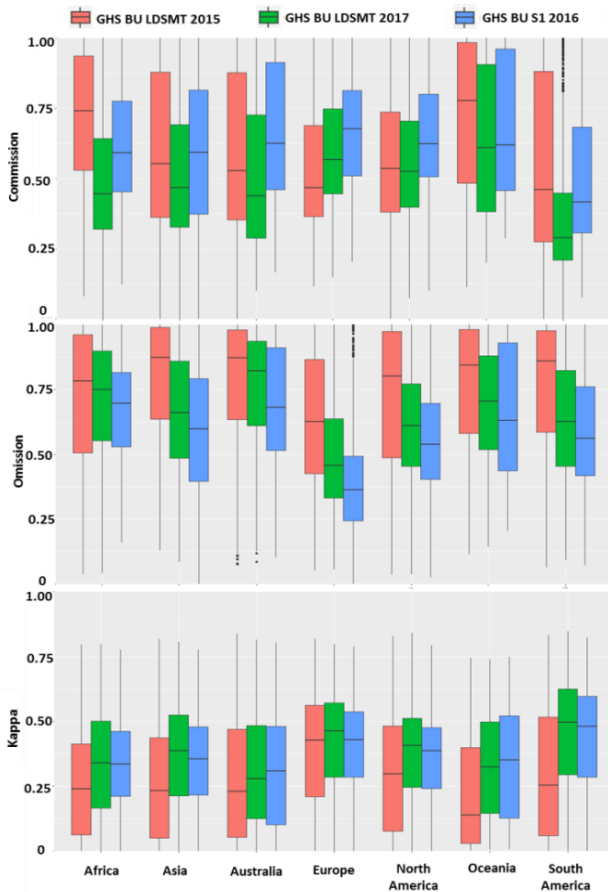


Fig.2 Accuracy assessment of built-up areas derived from Landsat (GHS BU LDSMT v.2015, GHS BU LDSMT v.2017) and Sentinel-1 (GHS BU S1 2016) using GUF as a reference.

The most noticeable improvements are observed in Africa where both GHS BU S1 2016, and to a higher extent, GHS BU LDSMT v.2017 contribute to the reduction of commission errors. In terms of omissions, Asia is the

continent where the highest misdetections were observed in the GHS BU LDSMT v.2015 product and where, thanks to the workflows run in the JEOPDD with Sentinel-1 and Landsat data, it was possible to significantly reduce those errors.

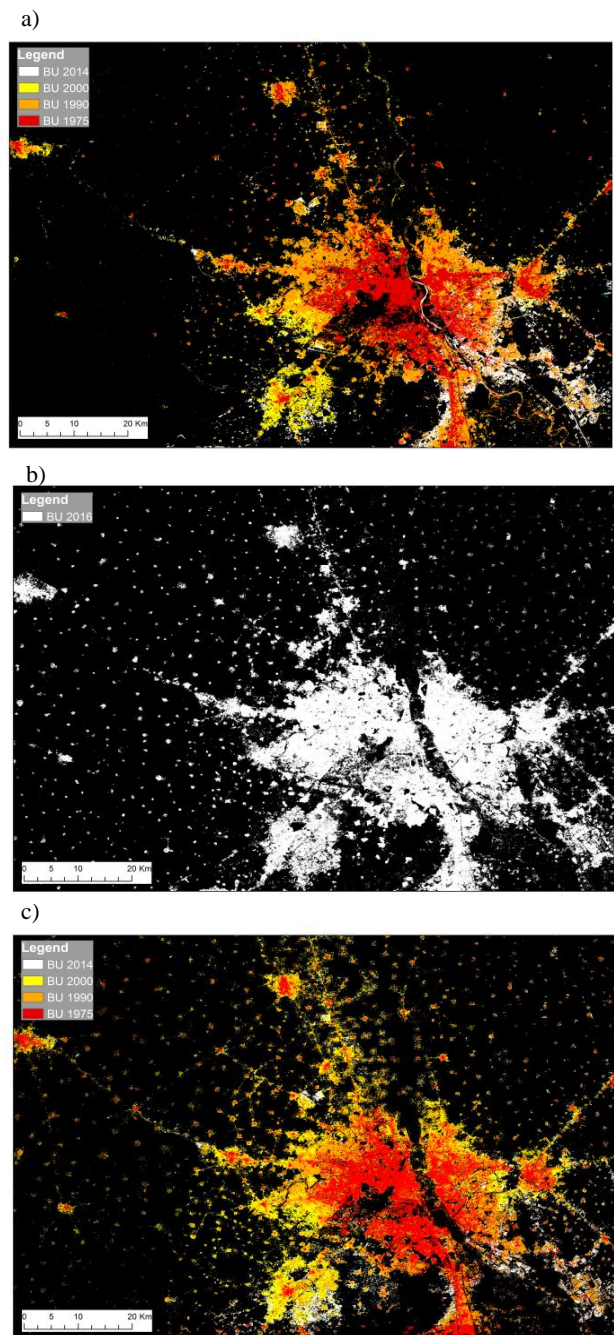


Fig.3 Example of built-up areas in New Delhi and surrounding area observed in a) GHS BU LDSMT v.2015, b) GHS BU S1 2016 and c) GHS LDSMT v.2017.

FIG.3b is an example of such enhanced capabilities covering the megacity of New Delhi and the surrounding villages. It corresponds to a typical case of under-detection of scattered settlements in rural areas in the first Landsat product (FIG.3a). The GHS BU LDSMT v.2017 which builds on the output from Sentinel-1 for the training of the SML also reflects the improvements, which propagate to the multitemporal output (FIG.3c).

The lowest median commission rate (0.27) is observed in South America with the GHS BU LDSMT v.2017 product whereas the lowest median omission rate (0.35) is achieved with the GHS BU S1 2016 product in Europe. The continent exhibiting the highest overall agreement with the GUF dataset is South America with a median Kappa coefficient of 0.50 obtained in the GHS BU LDSMT v.2017 product. It is followed by Europe and North America (with median kappa values of 0.42 and 0.40 respectively). These results give clear evidence that the incremental learning of the SML contributes to the reduction of regional differences in the performances as we notice a progressive stabilization of the Kappa values with the GHS BU LDSMT v.2017 product.

5. CONCLUSION

The true value of big earth observation data can be realized in a real remote sensing application [11]. Global mapping and monitoring of human settlements are an illustration of the challenges in managing, processing, and efficient exploitation of big earth observation data. In this paper, the latest developments in terms of processing big data for the purpose of improving the GHSL, firstly established with Landsat data, were presented. Two workflows with Sentinel-1 and Landsat data were run leveraging on the JRC JEODPP.

The latest advances in the production of GHSL data, building on previously acquired knowledge through incremental learning of the SML, are a strong experimental proof of the benefits of an integrated assessment of Landsat, Sentinel-1 and Sentinel-2 in support to the GHSL production. In this perspective the most efficient approaches for data/information fusion are being explored in the adaptive datacubes with a focus on the use of the hyper-temporal coverage of Sentinel-1 (A and B) and Sentinel-2 (A and B) that allow the computation of multi-temporal metrics. These metrics have the potential to further improve the delineation of built-up areas as demonstrated in previous studies [12]. The development of automated approaches for continual updating and for generating long-time series of built-up layers will be beneficial to applications that involve built-up gridded data as backbone. Such an application is the population disaggregation, which includes the GHSL built-up grid in the model covariates [13]. The scientific and technical challenges associated with these perspectives

provide the focus of future research on the global scale mapping of human settlements from space.

6. REFERENCES

- [1] T. Esch *et al.*, "Urban Footprint Processor--Fully Automated Processing Chain Generating Settlement Masks From Global Data of the TanDEM-X Mission," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, pp. 1617–1621, 2013.
- [2] J. Chen *et al.*, "Global land cover mapping at 30m resolution: A POK-based operational approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 103, pp. 7–27, May 2015.
- [3] M. Pesaresi *et al.*, "Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014," European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen, 2016.
- [4] M. Pesaresi, V. Syrris, and A. Julea, "A New Method for Earth Observation Data Analytics Based on Symbolic Machine Learning," *Remote Sens.*, vol. 8, no. 5, p. 399, May 2016.
- [5] P. Soille, A. Burger, D. Rodriguez, V. Syrris, and V. Vasilev, "Towards a JRC earth observation data and processing platform," in *Proceedings of the Conference on Big Data from Space (BiDS'16)*, Santa Cruz de Tenerife, Spain, 2016, pp. 15–17.
- [6] P. Soille *et al.*, "JEODPP: The JRC Earth Observation Data and Processing Platform," in *Proc. of the BiDS'17*, 2017, in this volume.
- [7] C. Corbane *et al.*, "Enhanced automatic detection of human settlements using Sentinel-1 interferometric coherence," *Int. J. Remote Sens.*, vol. in press. 10.1080/01431161.2017.1392642
- [8] V. Syrris, P. Soille, and C. Corbane, "A global mosaic from Copernicus Sentinel-1 data," in *Proc. of the 2016 conference on Big Data from Space (BiDS'17)*, Toulouse, France, 2017, in this volume.
- [9] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sens. Environ.*, vol. 37, no. 1, pp. 35–46, 1991.
- [10] R. G. Congalton, *Assessing the accuracy of remotely sensed data: principles and practices*, 2nd ed. Boca Raton: CRC Press/Taylor & Francis, 2009.
- [11] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proc. IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.
- [12] A. Lefebvre, C. Sannier, and T. Corpetti, "Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree," *Remote Sens.*, vol. 8, no. 7, p. 606, Jul. 2016.
- [13] S. Freire, D. Ehrlich, and S. Ferri, "Assessing Temporal Changes in Global Population Exposure and Impacts from Earthquakes," presented at the 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM), 2014.

ON THE CONTRIBUTION OF 20 YEARS OF ATSR DATA AND GEODESIC P-SPLINE EFFICIENT SPATIAL SMOOTHING METHOD TO ITCZ TREND ANALYSIS

E. Castelli¹, M. Ventrucchi², F. Greco², M. Valeri³, B.M. Dinelli¹, E. Papandrea⁴, S. Casadio⁴

¹ Istituto di Scienze dell'Atmosfera e del Clima (ISAC) CNR, Bologna, Italy

² Dipartimento di Scienze Statistiche "Paolo Fortunati", Università di Bologna, Bologna, Italy

³ Dipartimento di Fisica e Astronomia, Università di Bologna, Italy

⁴ Serco s.p.a., Frascati, Italy

ABSTRACT

The Intertropical Convergence Zone (ITCZ) is the region where the trade winds converge. It plays a key role in the general circulation of the atmosphere and understanding its variability is essential for improving global climate models.

We use the Total Column Water Vapour (TCWV) global distribution and the complementary cloud occurrence information to investigate the meridional migration of the ITCZ over 20 years from 1991 to 2012. The TCWV dataset was obtained from the analysis of the thermal infrared measurements acquired by the Along Track Scanning Radiometer (ATSR) instruments in the frame of the Long Term Data Preservation ESA programme. The huge amount of data used for the ITCZ analysis required the application of a method to efficiently extract the information. We use a Geodesic P-spline spatial smoothing method and the posterior probability distribution for identification of the ITCZ. Here we show the results of this work and the comparison with independent ITCZ datasets.

1. INTRODUCTION

The ITCZ region is characterized by strong convection. This results in heavy precipitation and high cloudiness. The ITCZ moves meridionally following the sun exposure, thus its position varies during the year. Its movements strongly affect the human life in the tropical region due to floods and drought periods related to ITCZ position. For this reason monitoring the ITCZ latitudinal displacement and its variation is extremely important. Furthermore, investigating the trend of this migration over the decades is of crucial interest. The ITCZ can be identified through the use of satellite data: outgoing longwave radiation (OLR) or seasonal mean precipitations are commonly used to identify the ITCZ. Other fields used for the ITCZ identification are wind fields and vorticity.

Moreover the global distribution of vertically integrated water vapour amount can be used for this purpose. The study presented here aims at the investigation of ITCZ meridional migration over the 1991-2012 period through the use of satellite data and a spatial smoothing technique. The data and

methods used for this work are briefly presented in section 2 and 3 while results are reported in section 4. Conclusions are given in section 5.

2. DATASETS

The ATSR instrument series onboard polar satellites from 1991 to 2012 [1] (ATSR-1 on ERS-1, ATSR-2 on ERS-2 and AATSR on ENVISAT) had as main objective the provision of Sea Surface Temperature (SST) data with high levels of accuracy and stability on global scale. These requirements are necessary for monitoring and carrying out research into the behaviour of the Earth's climate.

The ESA Earth Observation Long Term Data Preservation (LTDP) Programme aims at guarantee the preservation of the data from all EO ESA and Third Parties ESA managed missions on the long term. In the frame of this program, the ATSR Long Term Stability (ALTS) Project was designed to explore the key characteristics of the ATSR data set and new and innovative ways of enhancing and exploiting it (<https://earth.esa.int/web/sppa/activities/multi-sensors-timeseries/alts/>). One of the main outcome of this project was the development of the Advanced Infra-Red Water Vapour Estimator (AIRWAVE) methodology for the retrieval of TCWV from ATSR clear-sky over-sea Infra-red Brightness Temperatures at 11 and 12 micron [2] and the production of a dataset of Total Column Water Vapour (TCWV).

The TCWV dataset is produced exploiting the AIRWAVE-PP algorithm based on Python 2.7.6 with netCDF v.4.5.3, CODA v.2.11 and HDF5 v.1.8.14 libraries enabled. AIRWAVE-PP has been integrated into ESA Grid Processing on Demand (G-POD) [3]. The input dataset is made of about 20 years of data with more than 80000 orbits, for the total amount of more than 50 TB of data. The TCWV dataset is provided at both native spatial resolution (1x1 km²) and aggregated on a latitude/longitude grid (0.25x0.25°) for comparison purposes.

In this work we use both the TCWV dataset and a second one reporting the information on cloud frequency from ATSR data at coarse spatial resolution. The two datasets, whose to-

tal size is 50 GB, are complementary, can be used to extract information on ITCZ position on both land and sea, and are suitable for long term data analysis.

3. DATA ANALYSIS

The analysis of large datasets requires the adoption of techniques for the extraction of underlying information.

There are two main difficulties when building statistical models for spatial data collected at a global scale at a huge number of locations. First, the classic statistical models are computationally unfeasible, because model estimation requires the inversion of large dense matrices. Second, modelling spatial measurements collected worldwide arises issues on how to build valid covariance functions accounting for geodesic (arc-length) distances between data locations. To overcome this difficulty, in this work we undertake an efficient non-parametric approach that we dub *Geodesic P-splines* as it is an extension of the classic P-splines [4]. Classic Bayesian P-splines allow to obtain a smooth surface as a linear combination of bivariate B-splines basis functions, constructed by taking equally-spaced knots over the latitude-longitude domain, imposing smoothing constraints on the spline coefficients. From a computational point of view, this approach has several advantages due to the sparseness of the matrices involved in computation.

For these reasons, this model can be efficiently estimated using standard algebra for sparse matrices implemented in several statistical software. This model is very efficient and works well for data located over limited region of Earth, where Euclidean distances provide a reasonable approximation of the arc-length distances. In large regions, working with the euclidean distances will certainly lead to biased estimates. A straightforward approach to adapt classic P-splines for smoothing global data is to build the basis of equally-spaced B-splines directly over the sphere, assuming the sphere as an approximated representation of the Earth. We build this basis on a Geodesic Discrete Global Grid (GDGG) system [5].

The code used for the statistical analysis is developed using the R-INLA package [6]. This method is applied to ATSR TCWV and cloud occurrence data for each month to track the maximum of the convection activity. The algorithm to locate the ITCZ region is built by scanning the earth surface longitude-wise. At each longitude, we sample from the posterior predictive distribution at a fine grid over latitude. This allows to compute the posterior probability that a point at a given latitude belongs to the strip where the TCWV shows highest values. During the analysis we apply the method to both the monthly TCWV and cloud frequency datasets separately and to the merged dataset.

We processed the ATSR datasets (about 260 monthly files) on an Intel Xeon CPU E5-2637 v3 @3.5 GHz. Exploiting all the CPUs, the time elapsed for the calculations of one

month of data is about 2 hours, while the total computing time is 24 hours (time elapsed if no parallelization is used). This process produced a dataset of ITCZ probability distribution for each month, the size of one of these files is 16MB and the total size of the final ITCZ position archive is about 4GB.

4. RESULTS

In order to determine the performance of ATSR datasets in detecting the ITCZ position and set up the best configuration, we compare our results with the ones obtained using as input to the R-INLA code the TCWV GlobVapour dataset [7]. This dataset is composed by Special Sensor Microwave Imager (SSM/I) TCWV over the sea and Medium Resolution Imaging Spectrometer (MERIS) TCWV over land and has a spatial resolution of $0.5 \times 0.5^\circ$.

In Fig.1 we report, as an example, the ITCZ probability distribution (in red) obtained with ATSR TCWV (a), ATSR cloud occurrence (b) and merged datasets (c) in comparison with the results obtained with the GlobVapour TCWV dataset (d) for February 2003.

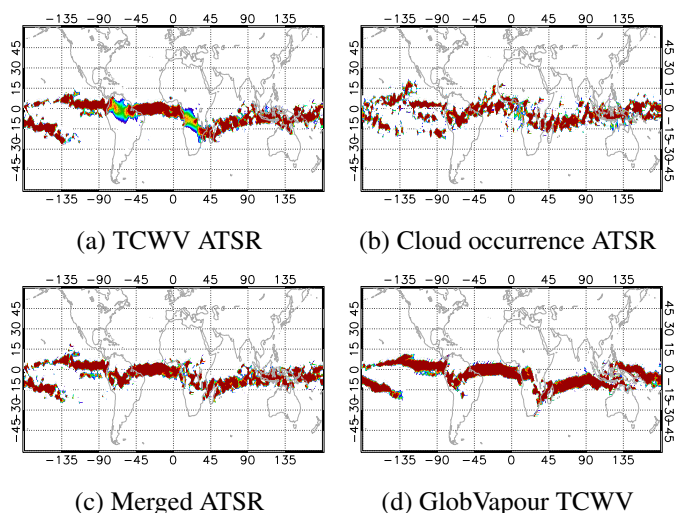


Fig. 1. ITCZ position in red for February 2003 from (a) ATSR TCWV, (b) Cloud occurrence ATSR, (c) merged and (d) GlobVapour TCWV datasets.

In Fig.2 (a) the median ITCZ position from the different datasets is reported while in (b) we report the differences between the ATSR datasets and the GlobVapour one. As can be noticed, the comparisons highlight the good performances of all the datasets with better results obtained for the merged dataset. Actually we get an average difference in ITCZ central position of $1.9 \pm 7.0^\circ$ for the Cloud dataset, $-0.6 \pm 6.6^\circ$ for TCWV dataset and $0.4 \pm 5.5^\circ$ for the merged one with respect to the GlobVapour dataset in February 2003.

This comparison highlight that the ATSR datasets contain valuable information for the investigation of the meridional position of ITCZ. The monthly results obtained with the

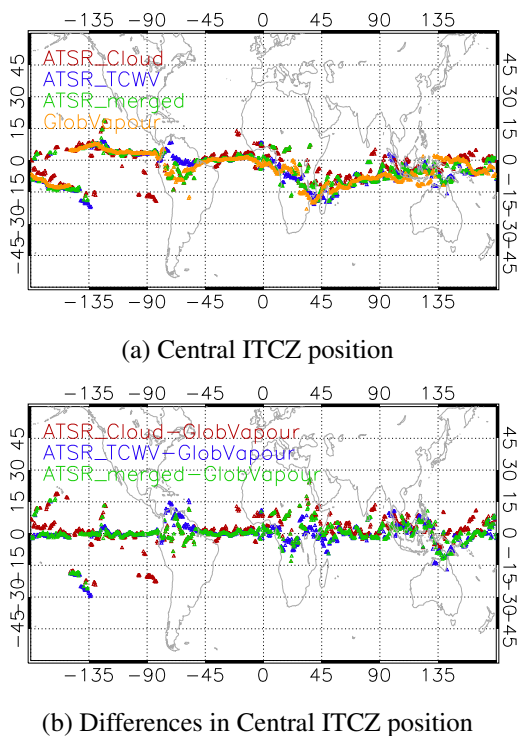


Fig. 2. (a) Central ITCZ position for February 2003, ATSR TCWV dataset (blue) ATSR cloud occurrence (red), ATSR merged (green) and GlobVapour dataset (yellow). (b) Differences between central ITCZ position in ATSR TCWV dataset (blue), ATSR Cloud occurrence dataset (red), ATSR merged (green) versus GlobVapour dataset for February 2003.

merged dataset are then used to investigate the trends, as a function of longitude, of the ITCZ meridional displacement through the use of a least-squares regression model that accounts for both seasonal and inter-annual components.

The ITCZ climatology and trends obtained in this analysis will then be compared to the ones obtained with the Global Precipitation Climatology Project (GPCP [8]) precipitation dataset over the same period. For the comparison we identify the ITCZ from the GPCP v2.3 CDR monthly dataset (spatial resolution of $2.5 \times 2.5^\circ$) as the region between $\pm 30^\circ$ latitude where the precipitation is maximum.

5. CONCLUSIONS

This work will show the valuable contribution to ITCZ trend study of long term data series of TCWV and cloud occurrence data obtained from the ATSR instruments. The adoption of a Geodesic p-spline smoothing statistical method specifically developed to extract information from large datasets is crucial for this analysis.

6. ACKNOWLEDGEMENT

Part of this work has been performed under the ALTS-LTDP project and under the ESA-ESRIN Contract No. 4000108531/13/I-NB.

7. REFERENCES

- [1] Delderfield, J., Llewellyn-Jones, D.T., Bernard, R., de Javel, Y., Williamson, E.J., Mason, I., Pick, D.R. and Barton, I.J., "The Along Track Scanning Radiometer (ATSR) for ERS-1", *Proceedings of SPIE 589*, pp. 114-120, 1986.
- [2] Casadio, S., Castelli, E., Papandrea, E., Dinelli, B. M., Pisacane, G., and Bojkov, B., "Total column water vapour from along track scanning radiometer series using thermal infrared dual view ocean cloud free measurements: The Advanced Infra-Red Water Vapour Estimator (AIRWAVE) algorithm", *Remote Sensing of Environment*, 172, pp. 1-14, 2016.
- [3] Burini, A., Casadio, S., Bojkov, B. R., Dinelli, B. M., Castelli E., and Papandrea eE, "The Advanced Infra-Red Water Vapour Estimator Prototype Processor (AIRWAVE-PP): tool design and ATSR processing", *SENTINEL-3 for Science 2015 Workshop*, Venice, Italy, 02-05 June 2015, 2015.
- [4] Eilers, P. and Marx, B., "Flexible smoothing with B-splines and penalties" *Statistical Science*, 11, pp. 89-121. 1996.
- [5] Sahr, K., White, D., and Kimerling, A. J., "Geodesic discrete global grid systems", *Cartography and Geographic Information Science*, 30(2), pp. 121-134, 2003.
- [6] Rue, H., Martino, S., and Chopin, N., "Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion)", *Journal of the Royal Statistical Society, Series B*, 71(2), pp. 319-392, 2009.
- [7] Lindstrot, R., Stengel, M., Schrder, M. Fischer, J., Preusker, R., Schneider, N., Steenbergen, T. and Bojkov, B. R., "A global climatology of total columnar water vapour from SSM/I and MERIS", *Earth Syst. Sci. Data*, 6, 221?233, 2014.
- [8] Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., et al., "The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979-present)", *Journal of hydrometeorology*, 4, pp. 1147-1167, 2003.

QA4ECV: 35 YEARS OF DAILY ALBEDO BASED ON AVHRR AND GEO

Said Kharbouche¹, Jan-Peter Muller¹, Olaf Danne², Nadine Gobron³

¹Mullard Space Science Laboratory, UCL, Holmbury St. Mary, RH5 6NT, UK

²Brockmann Consult GmbH, Max-Planck-Strae 2, D-21502 Geesthacht, Germany

³European Commission, Joint Research Centre, Directorate D - Sustainable Resources, Ispra, Italy

ABSTRACT

One of the objectives of the EU-FP7 project QA4ECV (Quality Assurance for Essential Climate variables, www.qa4ecv.eu, under contract N° 607405) project is to produce a long and consistent data record of global surface albedo. Thus, level-1 and level-2 data from several sensors, namely NOAA-AVHRR, several geostationary satellites and MODIS have been collected and pre-processed to ingest into our optimal estimation algorithm. Our primary output product is a daily global map of two types of albedo bi-hemispherical diffuse reflectance known colloquially as white sky albedo and direct hemispherical reflectance known as black sky albedo over three broadbands (vis: $0.4 - 0.7\mu m$; nir: $0.7 - 3\mu m$; sw: $0.4 - 3\mu m$) and, with a spatial resolution of $0.05^\circ \times 0.05^\circ$. From this product, we derive other up-scaled products such as monthly and $0.5^\circ \times 0.5^\circ$. The inter-comparison against third party products and in-situ data show a very good overall agreement and confirm the accuracy and the consistency of our output albedos.

Index Terms— Surface albedos, AVHRR, GEOS

1. INTRODUCTION

Surface albedo is one of the key parameters in global climate models and climate change studies. It can be used directly or indirectly via its derived essential climate variables such as FAPAR (Fraction of Absorbed Photosynthetically Active Radiation) and effective LAI (Leaf Area Index) which are also vital for climate modelling and biodiversity studies.

However, these global climate studies require several decades of albedo and often deploy ground-based albedo rather than remotely sensed albedo, which could increase the overall uncertainty because of the limitation of the spatial

coverage of these ground-based albedos. The reason why multi-decadal satellite-derived albedos are not heavily used is that these data do not fit requirements in terms of accuracy, type (white/black sky albedo), and also in terms of associated uncertainty and consistency.

Up until now, only the CLARA-A2 [1] provides the longest data record of satellite-derived albedo, these data covering the entire time period between 1982 and 2015 (34 years) globally with a spatial grid of $0.25^\circ \times 0.25^\circ$ and with two temporal resolutions: pentad (5-daily) and monthly. However, CLARA-A2 has two main limitations: it contains only one type of albedo which is the black sky albedo (DHR), and it covers only one spectral broadband with shortwave [$0.25\mu m, 2.5\mu m$]. These two limitations do not allow the derivation of vegetation-related ECVs such as FAPAR and LAI.

Within the EU-FP7 project (contract number of 607405) entitled QA4ECV (Quality Assurance four Essential Climate Variables), a key goal is to produce the longest and most consistent data record of global surface albedo with uncertainty provided on a pixel-by-pixel basis. Thus, as inputs, we have collected surface reflectance data that cover the whole daylight globe between mid-1981 and early 2017 with a spatial resolution of $0.05^\circ \times 0.05^\circ$ ($\approx 5km \times 5km$) and daily temporal resolution. These normalised surface reflectance data, which are atmospherically corrected, were provided by NASA and then processed into Bidirectional Reflectance Factors (BRFs) at the time of the satellite overpass by our partner at JRC for AVHRR (Advanced Very High-Resolution Radiometer). In parallel, geostationary panchromatic data from METEOSAT (e.g. MVIRI and SEVIRI instruments onboard the European satellites) for most of the same time period and GOES and GMS for some of the same time period were provided by our partner at EUMETSAT.

The output product contains two types of albedos: DHR (Directional Hemispherical Reflectance) and BHR (BiHemispherical Reflectance). Note that DHR and BHR are frequently referred to as black-sky-albedo and white-sky-albedo, respectively. The initial output products are daily $0.05^\circ \times$

The authors thank Eric Vermote and his co-workers at NASA Goddard Space Flight Center and the University of Maryland for processing the normalised surface reflectance product from all AVHRR sensors from 1981-2017; Jrg Schultz, Alessio Lattanzio and Youva Aoun of EUMETSAT for processing and providing the daily BRFs from METEOSAT, GOES and GMS and STFC, NERC and NCEO for providing the big data facilities at the CEMS-CEDA-JASMIN complex at the Harwell space campus. Finally, the authors would like to thank P. Lewis (UCL Geography) for his helpful advice.

0.05° grids with uncertainties attached at the per-pixel level. To provide additional products for direct use in climate models, these initial output products are up-scaled spatially to $0.5^\circ \times 0.5^\circ$ and temporally to monthly. The spatial coverage is the entire landmass, whilst the timeframe starts around 1982 and ends at the end of 2016. This results in a continuous gap-free 35-year albedo data record. The DHR and BHR are provided over three spectral broadbands: visible ($0.4 - 0.7\mu m$), near-infrared ($0.7 - 3\mu m$) and shortwave ($0.4 - 3\mu m$). Furthermore, a standard error and a cross-product alpha term derived from the full uncertainty covariance matrix (3×3) is also provided in this product for each BHR and DHR measurement. These products use netCDF 4.0 (CF-compliant) and have metadata fully consistent with all European Space Agency (ESA) Climate Change Initiative (CCI) products.

The next section gives an overview of the main stages of our production, whilst the following section presents some results of our validation process, and we end this paper by conclusions in the last section.

2. PRODUCTION

The production of albedo adopts a similar approach to the ESA GlobAlbedo products [2], which is based on optimal estimation using a climatology dataset of albedos that have been derived from 17 years of MODIS albedo (MCD43A1/2, Collection 6, 2000-2016) over the same broadbands. The production process comprises six main stages as listed below.

Firstly, BRFs of AVHRR (see a summary of those inputs in Figure 1) are re-masked with four types of classes (land, cloud, snow, undefined) using a specific probabilistic approach that was developed for this purpose. A probabilistic approach is applied in which, for each AVHRR channel, we model and fuse not only the ability to identify class(es) but also the ability to deny class(es). The accuracy of these masks was assessed against some of the masks from MODIS and, the results showed a high agreement. Figure 2 shows an overall statistics of our inter-comparison against 17 years of daily masks of MOD09 (MODIS surface reflectance, collection 6). Note that the approach is generic in such a way that it can be easily deployed for any multispectral sensor; thus a separate publication will be written for this new approach.

Secondly, we optimize three quadratic functions that map AVHRR BRFs in their original two bands ($[0.56\mu m, 0.68\mu m]$ and $[0.725\mu m, 1\mu m]$) to the three broadbands (VIS, NIR, SW) of the albedo climatology (MODIS). The optimization was performed by a dataset of BRFs that were collected from near-simultaneous observations from MODIS and AVHRR. Figure 3 shows the resultant regression plots with their resulting transformation coefficients. Note that the GEOs BRFs

are only available over shortwave and they were already converted using a quadratic function [3].

Thirdly, BHR and DHR are computed for both snow and snow-free conditions using AVHRR and GEO BRFs and the MODIS albedo climatology. The reason why we did not mix the samples of snow with those of snow-free is that the albedo creation for snow requires a very different approach.

Fourthly, after the creation of BHR and DHR for snow and snow-free, we merged all the albedos into a single final albedo product. Note that, we also keep snow and snow-free separate for specific requests from the climate modelling community that require unmixed snow and snow-free albedos.

Fifthly, we upscale our albedo spatially from 0.05° to 0.5° , and temporally from daily to monthly. Thus, four final albedos (BHR and DHR) products for three broadband (vis, nir, sw) are produced by the end of those processes: ($0.05^\circ \times 0.05^\circ$, daily), ($0.05^\circ \times 0.05^\circ$, daily), ($0.5^\circ \times 0.5^\circ$, monthly) and ($0.5^\circ \times 0.5^\circ$, monthly). Figure 4 shows two samples of BHR at shortwave for the same day of the year but for two different years: 1988 and 2008.

3. VALIDATION

The validation process is currently in progress and a final validation report will be publicly available through our website (www.qa4ecv.eu) at the end of this project (March 2018). But we provide an overview of the validation process and a sample of our validation output.

In addition to ground-based albedo, our output albedo are compared to surface albedo of third-party products such as GA.005 of ESA GlobAlbedo, MCD43C3 (of MODIS collection 6) CLARA-A2 (of CM-SAF, which also based on AVHRR), and VGT/Proba-V's albedo.

BHR and DHR represent the ratio of the portion of upwelling energy to downwelling energy. However, whilst BHR and DHR deploy the same upwelling energy calculation which is the integration of reflected energy over all angles, they differ in downwelling energy calculations. For BHR, the downwelling energy is assumed to be equally generated over the upper-hemisphere, whilst in the case of DHR, that energy is assumed to be generated by a single point (the Sun) of the upper-hemisphere. Thus we can say that BHR refers more to a very cloudy condition where thick clouds play the role of an ideal diffuser and, DHR refers more to cloud-free condition where downwelling energy is coming from a single point which is the sun.

BHR and DHR are two extreme cases that are not totally

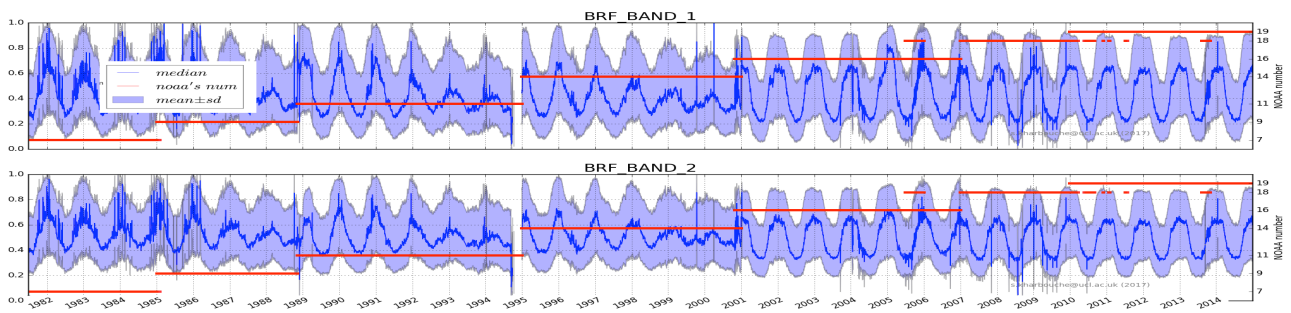


Fig. 1. time-series summary of unmasked AVHRR BRFs over the world's landmass that have been used as inputs for our albedo production. The red lines indicate each different NOAA-AVHRR. Note the difference of each sensor with corresponding jumps in reflectance

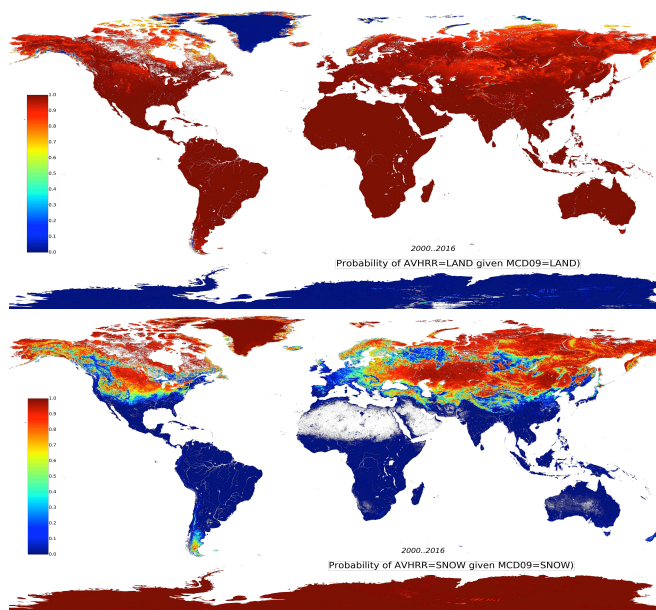


Fig. 2. Comparison of our cloud/snow/land masking against 17 years (2000-2016) of daily MOD09 masks (MODIS surface reflectance, collection 6). Top: the probability that AVHRR is land-flagged given MODIS is land-flagged. Bottom: the probability that AVHRR is snow-flagged given MODIS is snow-flagged.

realistic. The real measurable albedo that could be compared to ground-based albedometer measurements is a compromise between BHR and DHR, which is termed Blue-Sky-Albedo [4] and, can be calculated as follows:

$$\text{BlueSkyAlbedo} = \alpha \text{BHR} + (1 - \alpha) \text{DHR} \quad (1)$$

With $\alpha \in [0, 1]$ denotes the portion of diffuse component in downwelling energy during the time of satellite observation.

However, as we need to evaluate each albedo (BHR, DHR) separately, we adopted a new approach. Thus, to

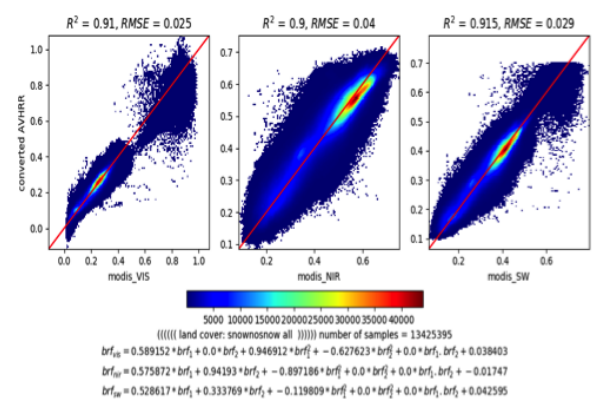


Fig. 3. More than 13 million (13.10^6) samples from near-simultaneous overpasses of AVHRR and MODIS were collected to optimize three quadratic functions that map AVHRR BRFs at their original two bands to the three broadbands of MODIS (VIS, NIR, SW).

evaluate BHR, we kept only ground data having a very high diffuse component ($\alpha > 0.98$), which refers to albedo measurements under very cloudy condition. Then we averaged these filtered data over a sliding time window of ± 8 days. For DHR, we kept only ground data having very low diffuse component ($\alpha < 0.1$) around local solar noon ($\pm 5^\circ$), which obviously refers to cloud-free conditions with low aerosol samples around that time of the day. Similarly, those collected data were also averaged using a sliding time window of ± 8 days. Thus, Figure 5 shows time series of shortwave albedos of our produced albedo side by side with those of the towers albedometer, GA and MODIS; over two SURFRAD sites (www.esrl.noaa.gov/gmd/grad/surfrad/): Sioux Falls, South Dakota, US, -SXF- ($lat = 43.73^\circ$, $lon = -96.62^\circ$), and Fort Peck, Montana, US, -FPK- ($lat = 48.30783^\circ$, $lon = -105.1017^\circ$), respectively.

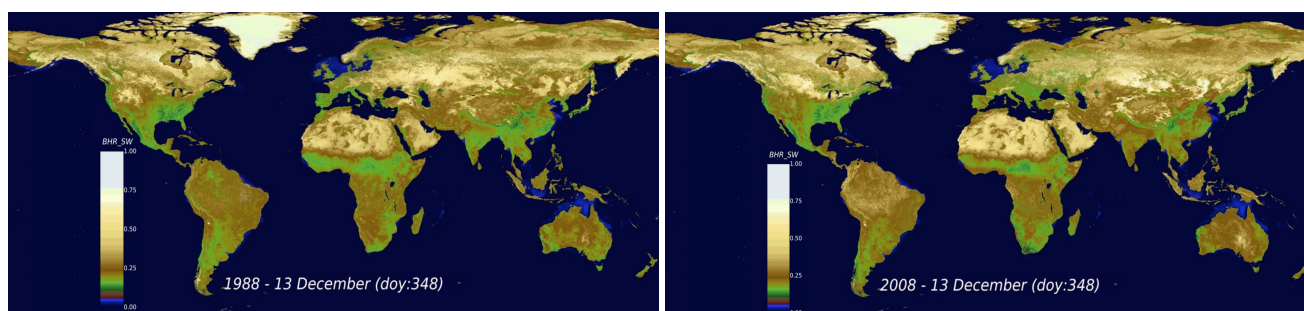


Fig. 4. A sample of our BHR shortwave products at $0.05^\circ \times 0.05^\circ$. Both images refers to the same day of year (13th December) but for different years: 1988 (left) and 2008 (right).

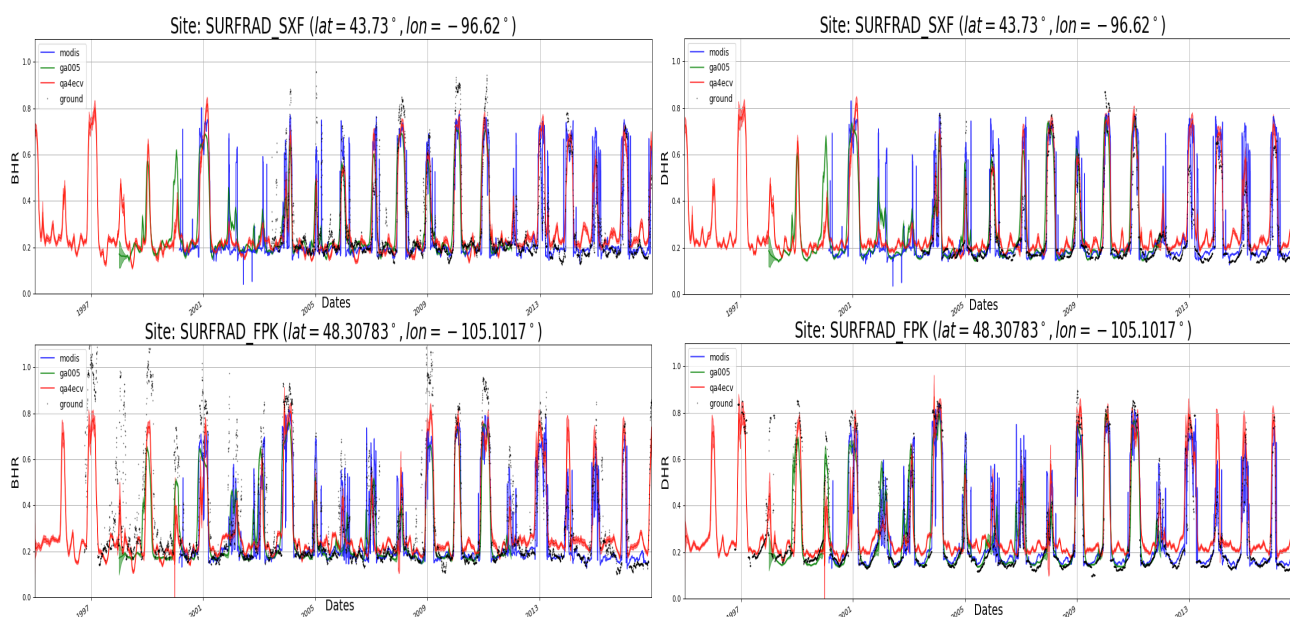


Fig. 5. side-by-side time series of BHR (left hand side figures) and DHR (right hand side figures) shortwave of GlobAlbedo, MODIS (collection 6), QA4ECV (AVHRR) and ground-based albedometer; over 2 SURFRAD sites: Sioux Falls, South Dakota, US (top), and Fort Peck, Montana, US (bottom).

4. CONCLUSIONS

The results obtained so far showed a good consistency despite the fact that the input data is very noisy and cloud contaminated, namely for GEO data. However, the associated pixel-based uncertainty is seen to be in good agreement with accuracy. QA4ECV albedo products will be unique in many aspects: long period, three broadband, spatial resolution, temporal resolution, accuracy, and associated uncertainty. Thus, we believe that this product will be a great benefit to climate modelling community.

5. REFERENCES

[1] K.-G. Karlsson et. al, "CLARA-A2: the second edition of the CM SAF cloud and radiation data record

from 34 years of global AVHRR data", "Atmospheric Chemistry and Physics. 17. 5809-5828." 10.5194/acp-17-5809-2017, 2017.

- [2] P. Lewis, J-P Muller et al. "GlobAlbedo Algorithm Theoretical Basis Document", <http://www.globalbedo.org>, 2011.
- [3] Y. Govaerts, et al. "Spectral conversion of surface albedo derived from Meteosat first generation observations." IEEE Geoscience and Remote Sensing Letters 3.1, pp23-27, 2006.
- [4] Schaeppman-Strub, Gabriela, et al. "Reflectance quantities in optical remote sensing Definitions and case studies." Remote sensing of environment, 103.1, pp 27-42, 2006.

EXPLORING VEGETATION PHENOLOGY AT CONTINENTAL SCALES: LINKING TEMPERATURE-BASED INDICES AND LAND SURFACE PHENOLOGICAL METRICS

R. Zurita-Milla *, R. Goncalves **, E. Izquierdo-Verdiguier *, F.O. Ostermann *

Faculty ITC - University of Twente * and NLeSC **, the Netherlands

ABSTRACT

Phenology is the science that studies the timings of recurring biological events such as leafing and blooming as well as their causes and variations in space and time. Spatially explicit environmental datasets and are key to understand phenological dynamics at continental to global scales. Here we present a novel exploratory analysis where we link temperature-based phenological indices and land surface phenological metrics derived from remotely sensed images. Our exploratory analysis, illustrated with two multi-decadal and high-spatial resolution phenological products for continental USA, focuses on identifying phenological regions and on mapping the coherence between phenological products. To cope with the computational challenges of analyzing big geo-datasets, we executed our analysis on a cloud platform running Apache Spark. First results show that weather, climate and land cover variability modulate phenological patterns in contrasting ways, and we believe that our computational solution work paves the path towards the analysis of global vegetation phenology at very high spatial resolution.

Index Terms— Extended spring indices, land surface phenology, exploratory data analysis, big geo-data, Apache Spark.

1. INTRODUCTION

Phenology studies the timing of recurring plant and animal biological phases, their causes, and their interrelations [1]. This seasonal timing varies from place to place and from year to year because it is strongly influenced by environmental conditions. Understanding this variability is critical to quantify the impact of climate change on our planet. In this work we present a novel exploratory analysis of two of the most important sources of spatio-temporal phenological data: phenological models based on weather- and location-related factors, and land surface phenological metrics derived from Earth observation sensors.

Phenological models. The Extended Spring Indices (SI-x; [2]) are a suite of models that transform daily temperatures into consistent phenological metrics that can be used to study the impact of global warming on vegetated canopies.¹ More

¹<http://www.globalchange.gov/explore/indicators>

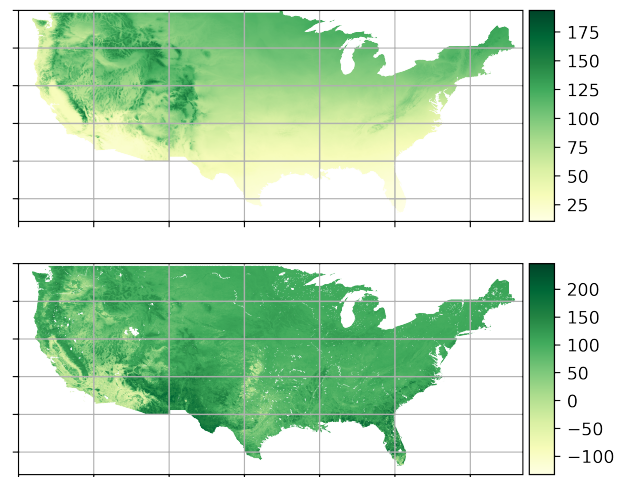


Fig. 1: Average of Leaf index [Top] and AVHRR SOS [Bottom] maps of contiguous North-America from 1989 to 2014.

precisely, the SI-x models predict the day of the year (DOY) of first leaf and of first bloom for three key indicator species [3]. These phenological dates can be used to track spring onset at specific locations by using data from weather stations [4] or at continental scales by using gridded weather and/or climatic datasets [5].

In this work, we use a new long-term (1980 to 2015) and high spatial resolution (1km) version of the Leaf and Bloom indices, which was recently generated for the coterminous US by adapting the SI-x models to a cloud computing environment [6]. Figure 1 [Top] illustrates, as an example, the average of the Leaf index from 1989 to 2014. This map shows a clearly noticeable spring gradient, with low values in the South and high DOY values in the North.

Land surface phenology. Time series of remotely sensed images can be used to derive various land surface phenological metrics. One of these metrics is the so-called Start of Season (SOS), which indicates the beginning of photosynthetic activity in plants. Several SOS products exist in literature. Often linked to a particular sensor or study. Here we use a SOS product specifically made for the US by processing

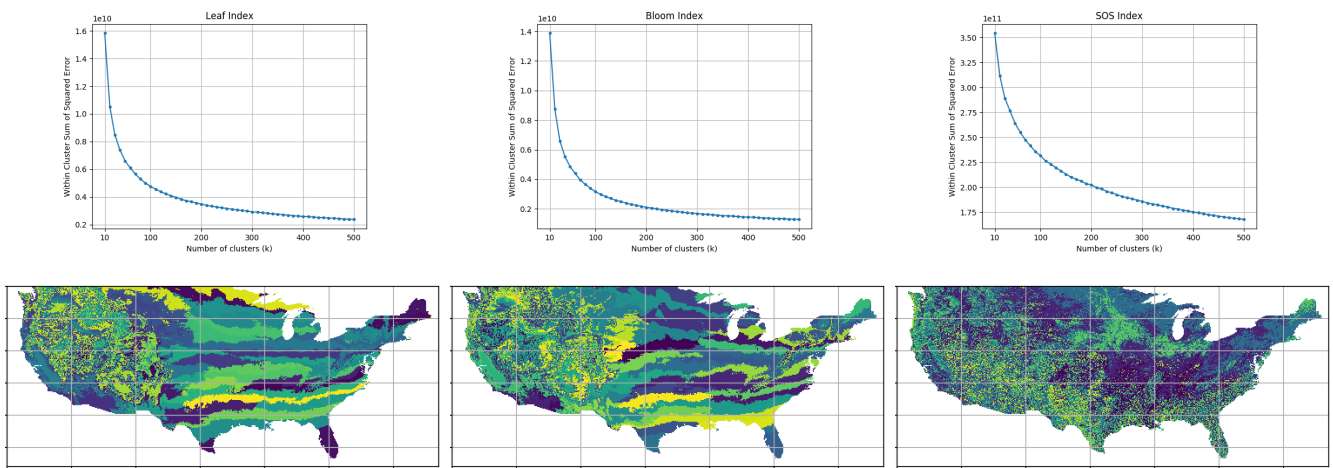


Fig. 2: Within cluster sum of squared Errors vs the number of clusters for the Leaf and Bloom indices and the SOS metric [Top row]. Clustering maps for the Leaf and bloom indices ($k=70$) and the SOS metric ($k=100$) [Bottom row]

time series of the Advanced Very High Resolution Radiometer (AVHRR) sensor². The AVHRR images were first transformed into a smooth time series of Normalized Difference Vegetation Index (NDVI). Then a curve derivative method was applied to predict NDVI values based on the previous observations. Finally, the SOS day was determined by identifying the day when the smoothed NDVI values become larger than the predicted NDVI values [7].

The spatial resolution of this product matches that of the SI-x but it is only available for the period 1989 - 2014³. Hence our exploratory analysis is based on the products available for this period. Again, as an example Figure 1 [Bottom] illustrates the average SOS values from 1989 to 2014. In this case the spring phenological gradient is less visible as the SOS depends on both the land cover and the weather conditions. Notice that the negative values in the SOS map indicate that the SOS took place the year before (i.e. in 1988).

Computational solution. Analyzing multi-decadal and very high spatial resolution phenological products at continental scales remains a challenging task. In this work we use a cloud-based solution based on Apache Spark [8] and its scalable machine learning library MLlib [9] to perform our exploratory data analysis. Given the lack of well-tested Spark solutions in the domain of big geo-data, a secondary aim of our work is to evaluate the potential of such a computational solution to analyze big raster datasets, in both local and cloud-based environments.

With the data stored in the original file formats, such as GeoTiff and HDF, users are able to analyze the data through Jupyter notebooks running either Python, R or Scala. These notebooks are not only used to share results among scientists but also as a provenance method for the scientific results.

²<https://lta.cr.usgs.gov/AVHRR>

³https://lta.cr.usgs.gov/avhrr_phen

Using the phenological products described above and our computational platform, we first identify regions with similar phenology (Section 2) and then study their correlation (Section 3). After that, we provide additional details on our computational platform (Section 4) and, finally, we summarize our findings and present follow up activities (Section 5).

2. MAPPING PHENOREGIONS

Clustering is a popular exploratory data analysis method that allows analysts to study their datasets at a higher level of abstraction [10]. Here we use K-means to identify regions with similar phenology (i.e. phenoregions). The three phenological products were clustered into k groups (with k values ranging from 10 to 500 in steps of 10) and the optimal k value was identified by the "elbow" of the Within Cluster Sum of Squared Error (WCSSE) graph. Figure 2 shows the WCSSE plots and the clustering results.

The optimal number of phenoregions is 70 for the Leaf and Bloom indices and 100 for the SOS metric. This indicates that land cover phenological variability is larger than the one caused by temperature differences. However, the phenological regions derived from the spring indices have a much stronger spatial coherence, especially on the East. Small scale differences in elevation and land cover lead to much more scattered phenoregions in the American West.

3. SPATIO-TEMPORAL CORRELATION

The ecological meaning of land surface phenological metrics is not fully clear yet [11]. To shed light on this, we performed a spatio-temporal correlation analysis between the Leaf and Bloom indices and the SOS metric. Figure 3 shows that large areas exhibit moderate to high positive correlations. This confirms that temperature is, indeed, one of the main drivers of

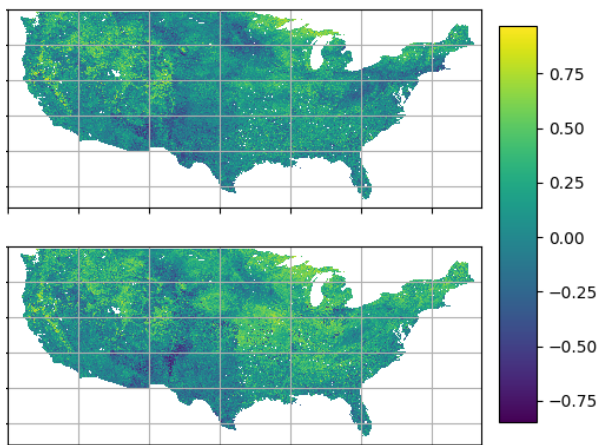


Fig. 3: Correlation between the Leaf index and SOS [Top] and between the Bloom index and SOS [Bottom]

phenological development. Our analysis also shows that the Leaf index is, in general, less correlated with the SOS than the Bloom index. This could indicate that satellites cannot detect the very early leaf onset, and that a certain amount of leaves (vegetation activity) is needed before spring can be seen from space.

Interestingly, Figure 3 also shows areas with moderate to high negative correlation. These areas correspond to locations where phenology seems to be driven by other environmental factors (e.g. water) and to areas where the SOS happens in the second half of the year.

4. COMPUTATIONAL PLATFORM

Our research work is conducted in an open-source platform using cloud-based infra-structures. With the aim either to do massive data analysis or a simple exploratory one, our computational platform is designed for easy user interaction and scalability. Users interact with the platform through Jupyter notebooks and computations are pushed down to a remote cluster. The computations are designed to use distributed data structures and Spark internals for efficient distributed processing. For its deployment and management we use Emma [12], a project to create a platform for development of applications for Spark and DockerSwarm clusters.

A cloud-based platform. The platform runs on an infrastructure composed by local or virtual machines attached to a large object storage with an Amazon Simple Storage Service (S3). The latter is becoming a de-facto API standard for objects-storage. It is supported by Google and Microsoft cloud services for easy port of cloud-based applications. The machines are prepared/constructed by either preparing cloud virtual machine or constructing using Vagrant [13] boxes. The



Fig. 4: Computational platform

latter allows the platform to be simulated on a local machine, i.e., provide a local development environment.

Once the machines are prepared the servers are provisioned using Ansible, an automation tool for IT infra-structure. Ansible [14] playbooks are used to create a storage layer, processing layer and JupyterHub [15] services. With Ansible we are able to deploy a platform with the same features at different locations, such as local cluster, national infra-structure or even a commercial cloud provider. Such feature allows us to have tool-provenance for easily repeatability of experiments between Scientists.

The platform's architecture is organized in three layers: storage layer, processing layer and JupyterHub services for user-interaction, (Figure 4). The storage layer offers two flavors of storage, file-base by Hadoop Distributed File System (HDFS), and object-based by Amazon S3 service. For local environments we use Minio [16], an open source object storage server with Amazon S3 compatible API, to avoid application re-write when moving to a cloud provider. HDFS is used by Apache Spark [8] to exploit data locality and to store intermediates to avoid re-computations. The object storage is used to store the phenology data products and other remote sensing data products.

At the processing layer we have Spark with its machine learning library SparkMLlib [9] and GeoTrellis [17] for high-performance geographic data processing. With GeoTrellis GeoTiffs are directly read from the S3 storage into Resilient Distributed Datasets (RDDs). With the phenology data products loaded as RDDs we then exploit Spark's internal for distributed data processing. One example is the mapping of pheno regions in Section 2.

For the data analysis the user expresses the operations either in Scala, R or Python using Jupyter notebooks. Hence, with a browser and remote connection the user is able to ex-

press a research question or collect an insight over large data sets. All computations are pushed down to the computational platform and results fetched back for data visualization.

Scalability. On our platform computations are not only pushed down for remote processing, but they are also designed to exploit Spark's cluster computational features. To achieve that data is always loaded into memory-based data structure such as RDD, DataFrames and distributed matrices. With the data loaded into Spark's memory-based structures, distributed task scheduling and fault-tolerance is then handled by Spark.

Such strategy is crucial to achieve efficiency and scalability. It also releases the user from the burden of re-writing an application in case the problem size increases, e.g., use higher resolution data from Sentinel-2, or for changes in the amount of available resources when moving to a different cloud-infrastructure. The decision of which structure to use and a study on the impact of different resource allocation, i.e., a detailed performance profile, is out of the scope of this paper.

5. CONCLUSIONS AND FUTURE WORK

In this paper we exploit the Apache Spark ecosystem for large scale distributed processing. With our phenological experiments we have demonstrated that it possible to map phenoregions at high spatial resolution and at continental scales. Moreover, we have shown that temperature-based indices are both positively and negatively correlated with the AVHRR SOS metric. Further analysis is needed to better understand the complementary and synergistic value of these two phenological products.

Future work will deal with the integration of the millions of ground phenological observations collected by citizen scientists as well as with the analysis of very high spatial resolution phenological metrics from the Sentinel missions. We plan to conduct this analysis at the ESA cloudtoolbox⁴ and at different commercial cloud providers in an attempt to verify if our platform is generic enough.

Acknowledgments

This work has been partially supported by the NLeSC Project: "High spatial resolution phenological modelling at continental scales"⁵ and it was carried out using the Dutch national e-infrastructure provided by the SURF Cooperative. The extended spring indices were computed in the framework of the "Green-wave" project, funded via a Google Faculty Award to the first author of this paper.

⁴<http://eogrid.esrin.esa.int/cloudtoolbox>

⁵<https://github.com/phenology>

6. REFERENCES

- [1] H. Lieth, "Purposes of a phenology book," in *Phenology and seasonality modeling*, 1974.
- [2] M. D. Schwartz, T. R. Ault, and J. L. Betancourt, "Spring onset variations and trends in the continental united states: past and regional assessment using temperature-based indices," *International Journal of Climatology*, 2013.
- [3] A. H. Rosemartin, E. G. Denny, J. F. Weltzin, R. L. Marsh, B. E. Wilson, H. Mehdipoor, R. Zurita-Milla, and M. D. Schwartz, "Lilac and honeysuckle phenology data 19562014," *Scientific Data*, vol. 2, 2015.
- [4] T. R. Ault, R. Zurita-Milla, and M. D. Schwartz, "A matlab@ toolbox for calculating spring indices from daily meteorological data," *Computers & Geosciences*, vol. 83, pp. 46 – 53, 2015.
- [5] E. Izquierdo-Verdiguier, R. Zurita-Milla, T. R. Ault, and M. D. Schwartz, "Development and analysis of spring plant phenology: 36 years of 1-km grids over the conterminous us," *Agricultural and Forest Meteorology*.
- [6] E. Izquierdo-Verdiguier, R. Zurita-Milla, T. R. Ault, and M. D. Schwartz, "Using cloud computing to study trends and patterns in the extended spring indices," *Third International Conference on Phenology*, 2015.
- [7] B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen, "Measuring phenological variability from satellite imagery," *Journal of Vegetation Science*, vol. 5, no. 5, pp. 703–714, 1994.
- [8] "Apache spark," <https://spark.apache.org>.
- [9] "Apache spark-mllib," <https://spark.apache.org/mllib>.
- [10] X. Wu, R. Zurita-Milla, and M. J. Kraak, "A novel analysis of spring phenological patterns over europe based on co-clustering," *Journal of Geophysical Research: Biogeosciences*, 2016.
- [11] M. A. White, de K. M. Beurs, K. Didan, D. W Inouye, A. D Richardson, O. P. Jensen, J. O'keefe, G. Zhang, R. R. Nemani, et al., "Intercomparison, interpretation, and assessment of spring phenology in north america estimated from remote sensing for 1982–2006," *Global Change Biology*, 2009.
- [12] R. Goncalves, S. Verhoeven, N. Drost, and J. Attema, "Emma," doi:10.5281/zenodo.996308.
- [13] "Vagrant," <https://www.vagrantup.com>.
- [14] "Ansible," <https://www.ansible.com>.
- [15] "Jupyterhub," <https://github.com/jupyterhub/jupyterhub>.
- [16] "Minio," <https://www.minio.io>.
- [17] "Geotrellis," <https://geotrellis.io>.

LARGE SCALE FLOOD RECURRENCE MAP USING SAR DATA

Marco Chini¹, Ramona Pelich¹, Renaud Hostache¹, Patrick Matgen¹,
Jose Manuel Delgado^{2,3}, Giovanni Sabatino^{2,3}

marco.chini@list.lu

¹Luxembourg Institute of Science and Technology (LIST),

Environmental Research and Innovation Department, Luxembourg

²Progressive Systems Srl, Parco Scientifico di Tor Vergata, 00133 Roma, Italy

³ESA Research and Service Support, via Galileo Galilei, 1, 00044 Frascati, Italy

ABSTRACT

A newly developed automatic method to map flooded areas, using SAR data, has been applied to the archive of Envisat ASAR mission in order to generate large scale flood recurrence maps. The flood mapping algorithm makes use of hierarchical image tiling, histogram thresholding and region growing to delineate the flood extent. We define the flood recurrence of a specific area as the pixel-based sum of all the reliable binary flood maps. SAR water-like surfaces such as shadow, tarmac and absorbing vegetation are filtered out using auxiliary data sources such as the height above nearest drainage index, land cover maps and a digital elevation model. In order to demonstrate the effectiveness of this methodology, the algorithm has been already applied to the Envisat ASAR images archive over the entire Europe and we are currently in the phase of applying it to the entire globe. The processing has been possible thanks to the ESA's Research and Service Support team, through the Grid Processing On Demand environment (G-POD).

Index Terms— Automatic flood extent delineation, flood recurrence, global maps, HAND-index, Envisat SAR archive.

1. INTRODUCTION

A large collection of Synthetic Aperture Radar (SAR) images is available from past (e.g. ERS, Envisat) and current satellite missions (e.g. Sentinel-1), providing an almost complete global coverage and enabling the generation of a world flood record. Many studies making use of optical data, such as Modis and Landsat, have assessed and developed global-scale flood maps [1-2], although optical images are sensitive to cloud coverage and dependent on daylight. SAR sensors occupy a privileged place in flood mapping applications because of the microwave radiation sensitivity to the presence of surface water. Moreover, they provide synoptic views and are

capable of imaging quasi all-weather day/night observations [3-4]. In SAR images, the high contrast between flooded and non-flooded terrain is due to the fact that smooth water surfaces reflect the incident radar signal in the specular direction. This generally results in markedly low backscatter value recordings. Moreover, the rough non-flooded terrain scatters the signal in many different directions, thereby producing much higher backscatter value recordings. There are a number of exceptions, such as shadow, wet snow or vegetated regions, which, under specific acquisition geometry have a similar backscatter to the one of water. This leads to misclassification in SAR-based flood mapping algorithms [5].

The most commonly applied SAR-based methods to map water/flood regions is thresholding, which is arguably the most rapid technique. The threshold is usually identified as the maximum backscattering value representatives of open water, which is a compromise between the minimization of over and under detection, and all pixels having an intensity value lower than this fixed threshold value are classified as inundated. Parametric algorithms are used to automatically extract an optimal threshold value but their efficiency is known to be strongly hampered when the respective fraction of the image occupied by the different classes is strongly unbalanced or if the distribution functions significantly overlap.

In this paper we propose to generate the first global database of flood inundation maps derived from large SAR data collections. In this context, we will firstly use a hierarchical split-based approach (HSBA) for parametric thresholding SAR images [6], in combination with a region growing approach which permits to take into account contextual information, to rapidly delineate flooded areas. Once the flood delineation algorithm is applied to an entire SAR image archive, the resulting binary flood maps are used to attribute a flood-recurrence score to each pixel. Pre- and post-processing steps are required in order to eliminate

the non-flooded regions that exhibit radar responses similar to the water surface. With this purpose we employ several external data sources such as Height Above Nearest Drainage (HAND) index [7], Local Incidence Angle (LIA) and land cover maps, such as the CORINE Land Cover (CLC) inventory [8]. In collaboration with ESA's Research and Service Support team, the consolidated flood mapping software has been integrated within the Grid Processing On Demand environment (G-POD) [9] for processing the entire archive of Envisat ASAR-WS and large SENTINEL-1 imagery collections.

The ESA Research and Service Support service (RSS) [9] has the mission to provide tools and services to support the EO community in exploiting data, researchers in developing new algorithms and applications, and service providers in generating and delivering value added information. The grid processing on-demand service (G-POD) provided by RSS, is based on an operational processing environment where specific algorithms provided by scientists can be integrated for processing and exploiting EO data from distributed archives.

2. METHOD AND RESULTS

The flood mapping methodology consists of an automatic histogram thresholding followed by a region growing process. As mentioned before, in SAR images, water areas are characterized by low backscatter values. From a statistical point of view, the histogram of water and non-water backscattering values is generally characterized by a bimodal distribution. Since flooded areas often represent a low fraction of SAR scenes, the parametrization of water/flood distribution from an entire SAR image might be not possible. Therefore, before parameterizing it, a split of the entire SAR image into several tiles is required. Here we utilize HSBA-Flood algorithm [6], which instead of fixing the tile size a priori makes use of a hierarchical splitting framework, where tiles with identifiable bimodal distributions, are automatically selected. The hierarchical tiling of the image is done using a quad-tree decomposition, which consists of iteratively decomposing image regions into four equally sized quadrants (i.e., tiles), the so-called sub-quadrants (an exhaustive description of the algorithm is in Chini et al. [6]). Based on the selected tiles, water/non-water classes distributions are estimated with the purpose to select the threshold for separating the water class from the rest. The most straightforward way for determining the threshold is to visually examine the histogram and to place the threshold in the relative minima between the two modes. Frequently, the two modes are not completely separated and overlap region might be present also when using HSBA, in spite of the capacity of the algorithm to estimate the two-class distributions. In order to handle this kind of situation, the selection of the threshold can benefit from the

combination of the contextual information of the image with its intensity information. In HSBA-Flood the contextual information is addressed using a region growing step assuming that pixels constituting the target class are clustered rather than randomly spread out over the entire image. The seeds of the region growing step are extracted based on the water distribution, i.e. all pixels with a backscattering value lower than the mean value of the water distribution. Within this step the seeds are iteratively grown testing many different tolerance thresholds. To find the optimal thresholds in order to stop the growing process, we minimize the root mean square error between the empirical distribution estimated from the region growing pixels and the theoretical distribution of the water pixels.

The objective of this work is to provide a flood record from the entire Envisat SAR archive. Thus in order to reduce the processing time and to overcome the cumbersome step of selecting the reference image, which is usually required in change detection-based flood mapping algorithms, here the HSBA-Flood algorithm is applied without change detection. In this case the algorithm search for the water class, i.e. the darkest class in the image, and based on its distribution classifies each image.

Ambiguities caused by several water-like surfaces or permanent water bodies are eliminated making use of the HAND index, LIA and CLC map. The ambiguities that have been considered are the following:

Shadow: The radar imaging geometry and surface topography cause radar shadow, which has backscatter values similar to the water ones. The LIA is a relevant parameter to detect shadowed regions. The LIA has been extracted making use of the SRTM DEM and then employed in the pre-processing step, permitting to eliminate shadow pixels from input SAR images.

Permanent water bodies: This class is masked out in the pre-processing step making use of land cover maps available, such as CLC, in order avoid false alarms provided by big river or lakes. It is worth to consider that the resolution of ASAR-WS is 150 m, thus only big rivers are concerned.

False alarms in mountainous areas: Supposing that flooding events occur contiguously of drainage systems, pixels classified as inundated by the flood mapping algorithm but located at a higher altitude with respect to the one of the drainage system can be considered as false alarms. To this aim the HAND index has been used to mask out all areas that based on hydraulic consideration cannot be inundated. HAND is defined as the elevation difference between a topographic reference and the height level of the nearest drainage system. This can help to remove false alarms caused wet snow, which is very absorbent and produces very low backscatter which can be easily misinterpreted as floodwater.

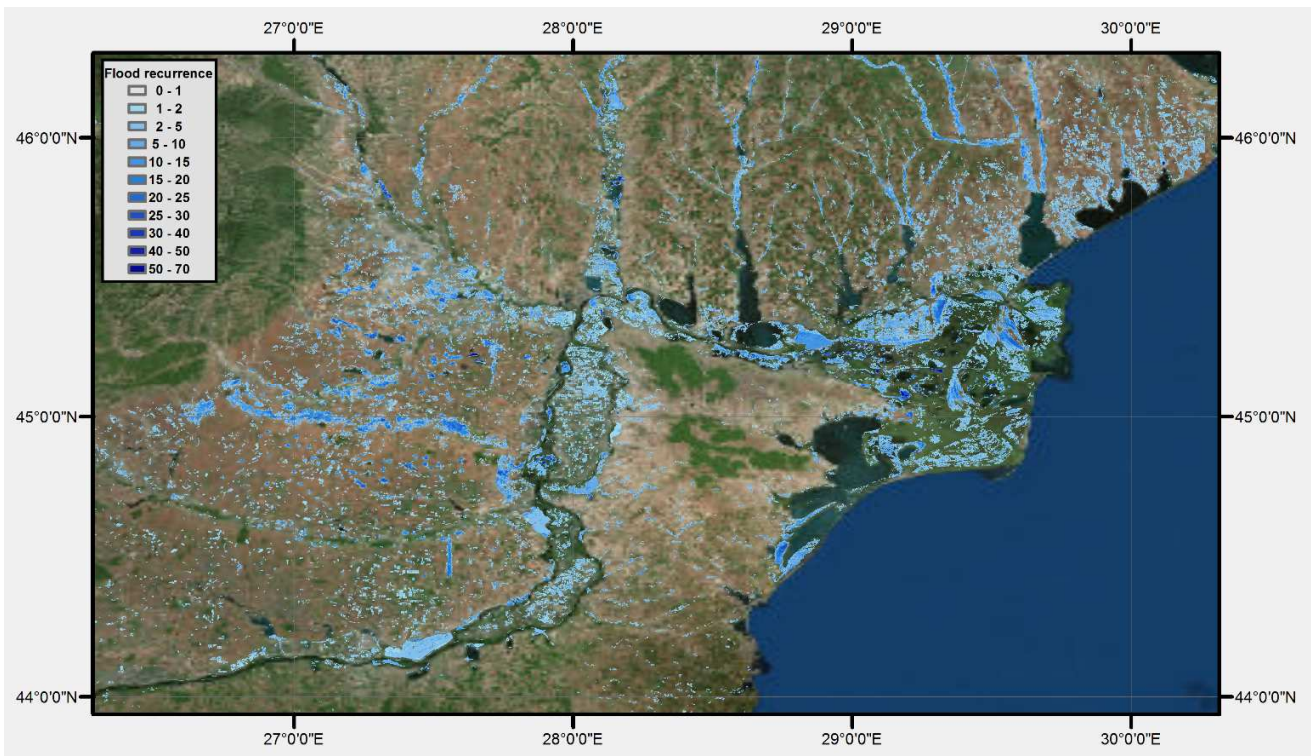


Figure 1: Flood recurrence map derived from Envisat ASAR images over Danube Delta.

Water-like surfaces: Also for this false alarm, land cover maps represent another auxiliary data source that are appropriate to identify this the flooding ambiguities. From the CLC map we identified the location of airport runways or certain urban areas presenting water-like SAR backscattering values.

An example of a flood recurrence map generated over the Danube Delta is shown in Figure 1. The HSBA flood mapping algorithm has been applied to a total of 900 Envisat ASAR images acquired over this area during the sensor's entire period of sensing from 2002 to 2012. After a post-processing operation based on different external sources as the HAND index, the final recurrence map is defined as the pixel-based sum of all the reliable binary flood maps. We notice from the resulted flood recurrence map that the identified flood areas are located near the main river streams, along the Danube's course and in proximity of Danube Delta wetlands.

Figure 2 gives another example of a flood recurrence in the region of the Po River in Italy. The flood recurrence map was obtained by processing about 700 Envisat ASAR images acquired from 2002 to 2012. This area near Vercelli, Italy is well known for the cultivation of rice. The majority of rice fields are typically flooded during certain planting and growing phases. This explains the high flood recurrence over large surfaces in this region. Rice fields are one of the water ambiguities radar signatures due to vegetation. A

possible solution in filtering such areas from SAR-based flood maps could be the use of Normalized Difference Vegetation Index (NDVI).

3. ACKNOWLEDGMENT

M. Chini, P. Matgen and R. Hostache work was supported by the National Research Fund of Luxembourg (FNR) through the MOSQUITO (C15/SR/10380137) project.

4. REFERENCES

- [1] G.R.Brakenridge, "Global active archive of large flood events," in *Dartmouth Flood Observatory*, University of Colorado.
- [2] B. Revilla-Romero, F. A. Hirpa, J. Thielen-del Pozo, P. Salamon, R. Brakenridge, F. Pappenberger, and T. De Groeve, "On the Use of Global Flood Forecasts and Satellite-Derived Inundation Maps for Flood Monitoring in Data-Sparse Regions," *Remote Sensing*, 7(11), 15702-15728, 2015.
- [3] L. Pulvirenti, M. Chini, and N. Pierdicca, "Use of SAR data for detecting floodwater in urban and agricultural areas: The role of the interferometric coherence," *IEEE Transactions on Geoscience and Remote Sensing*, 54 (3), 1532 – 1544, 2016.
- [4] L. Giustarini, R. Hostache, D. Kavetski, M. Chini, G. Corato, S. Schlaffer, and P. Matgen, "Probabilistic Flood Mapping Using Synthetic Aperture Radar Data," *IEEE Transactions on Geoscience and Remote Sensing*, 54, 6958-6969, 2016

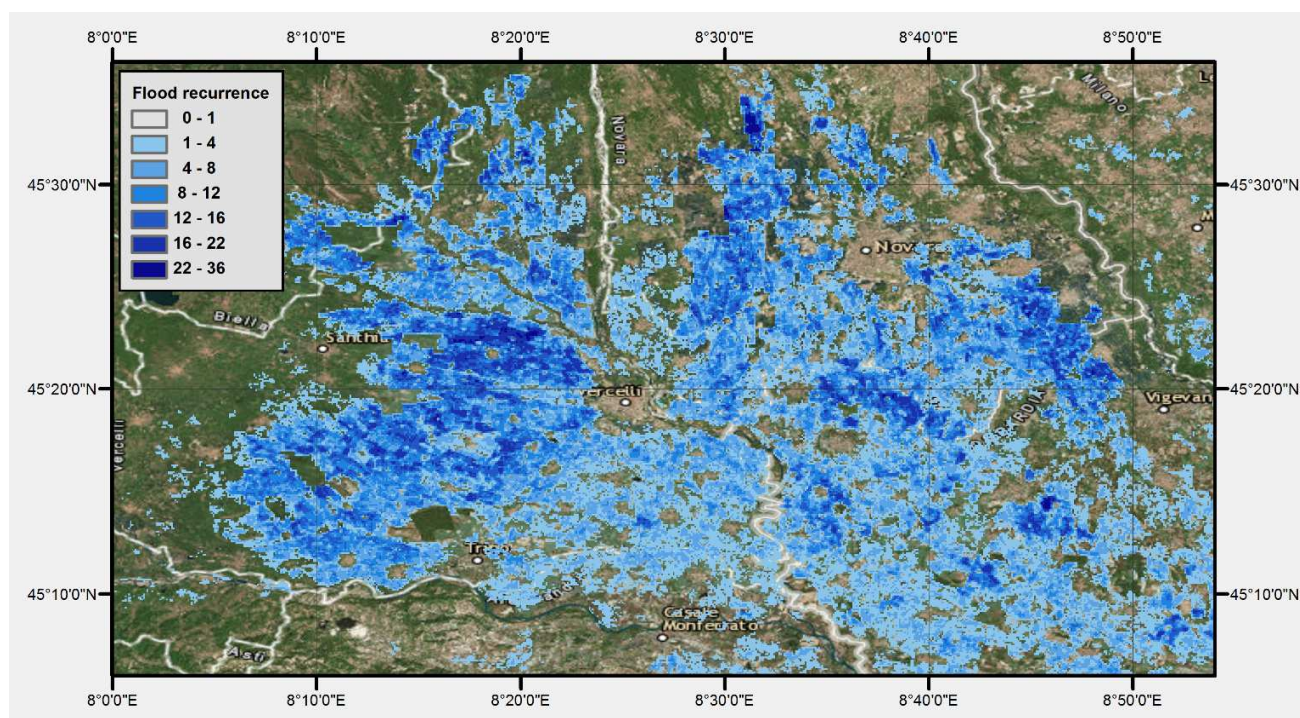


Figure 2: Flood recurrence map derived from Envisat ASAR images over the Po River region, Italy.

[5] L. Pulvirenti, F. S. Marzano, N. Pierdicca, S. Mori, and M. Chini, "Discrimination of water surfaces, heavy rainfall and wet snow using COSMO-SkyMed observations of severe weather events," *IEEE Transactions on Geoscience and Remote Sensing*, 52 (2), 858-869, 2014.

[6] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A Hierarchical Split-Based Approach for parametric thresholding of SAR images: flood inundation as a test case" *IEEE Transactions on Geoscience and Remote Sensing*, 10.1109/TGRS.2017.2737664, 2017.

[7] European Environment Agency, "Corine land cover (CLC) 2012, version 18.5.1," <http://land.copernicus.eu/pan-european/corine-landcover/clc-2012/view>.

[8] A.D. Nobre, L.A. Cuartas, M. Hodnett, C.D. Renno, G. Rodrigues, A. Silveira, M. Waterloo, and S. Saleska, "Height above the nearest drainage a hydrologically relevant new terrain model," *Journal of Hydrology*, 404 (1-2), 13-29, 2011.

[9] P.G. Marchetti, G. Rivolta, S. D'elia, J. Farres, N. Gobron, and G. Mason, "A model for the scientific exploitation of earth observation missions: The esa research and service support," *IEEE Geosci Newsl*, 162, 10-18, 2012.

INTERACTIVE VISUALISATION AND ANALYSIS OF GEOSPATIAL DATA WITH JUPYTER

D. De Marchi, A. Burger, P. Kempeneers, and P. Soille

European Commission, Joint Research Centre (JRC)

Directorate I. Competences, Unit I.3 Text Data Mining, via Fermi 2749, 21027 Ispra (VA), Italy

ABSTRACT

With its open-source policy and accommodation of a wide variety of programming languages, the Jupyter web-application has recently positioned itself as the most popular environment for interactive scientific computing. In this paper, the use of Jupyter notebooks based on IPython for interactive visualisation and analysis of geospatial data is put forward and used as front-end to a back-end platform with petabyte scale storage and processing capabilities. Deferred processing allows computations to be restricted to the zoom level and extent of the area displayed in a map viewer.

Index Terms— deferred processing, Sentinel, Copernicus, visualisation, Docker, Jupyter, IPython

1. INTRODUCTION

Web-based interactive computational environments have recently gained a lot of interest for data analysis in all scientific fields. This can be explained by the ease of use (no software besides a browser needs to be installed) and the possibility to have the server side co-located with data storage and processing capabilities. Among the numerous web-applications for data analysis, Jupyter [3] was chosen for its open-source policy, its wide user-basis in many scientific fields, and its flexibility to serve a range of programming languages. In addition, Jupyter notebooks provide a unique environment for integrating code, documentation, and publication in a single source file, thereby contributing to knowledge sharing and collaborative working.

The developments presented in this paper are implemented on the JRC Earth Observation Data and Processing Platform (JEODPP) [12]. This platform serves the needs of JRC policy support activities requiring big data capabilities for analysing geospatial data. The JEODPP can be viewed as a three layer pyramid with a petabyte scale storage and processing basis. The first layer accommodates massive batch processing. The second layer provides a remote desktop environment with all software needed for further developing legacy applications. Interactive visualisation and analysis is provided by the third layer (tip of the pyramid). The interactivity is enabled by a web-based environment integrated in a Jupyter notebook [3].

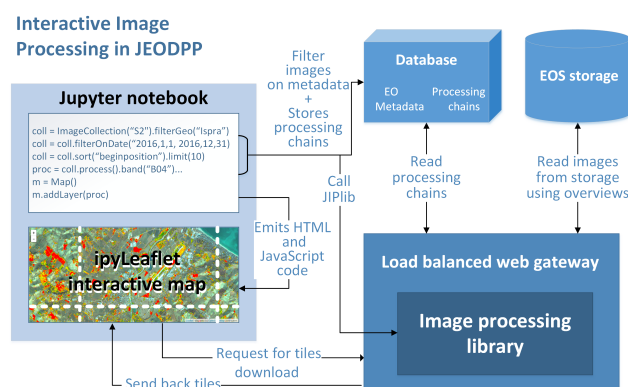


Fig. 1: The interactive processing and visualisation model.

2. INTERACTIVE ENVIRONMENT OVERVIEW

An overview of the proposed interactive processing operation mode integrated on the JEODPP is sketched in Fig. 1. The Jupyter notebook provides a programming interface that can accommodate a variety of programming languages. The Python language was selected for its open source and its wide variety of packages for data scientists with processing, analysis, and visualisation capabilities. The Python code developed in the Jupyter notebooks is not directly executing the data processing. Indeed, the processing is merely defined as a deferred execution pattern that is only executed when needed [9]. The code from the Jupyter notebook is translated into a JavaScript Object Notation (JSON) describing all processing and analysis logic that is matching the desired image processing chain object as well as the desired data on which it needs to be applied thanks to a selection process based on metadata information. A key element of the Jupyter notebook is the interactive map display. This map relies on the Leaflet JavaScript mapping library, loaded into Jupyter notebook via the IPyleaflet extension. The map contains a selectable base background layer for navigation. The deferred processing is executed when adding a processing chain as a display layer to the interactive map.

The OpenStreetMap defines the default base map for the map viewer. Other base maps such as OpenTopoMap, OpenMapSurfer or the MODIS global composite of any specific

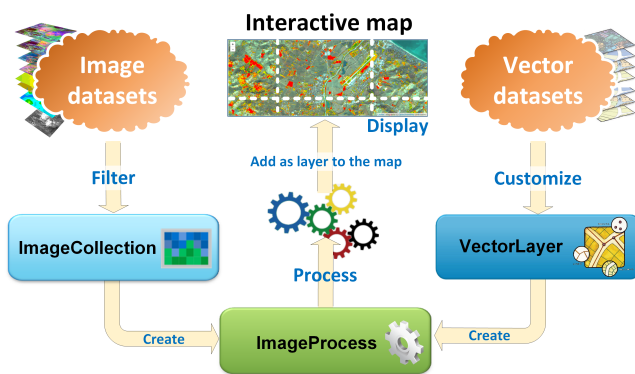


Fig. 2: The main components of the proposed Jupyter-based interactive environment for geospatial data visualisation and analysis [11].

date can be selected. Any given collection of images or vector layers can then be viewed on the top of the base maps while considering a user-defined opacity level. Available collections on the JEODPP platform are based on radar imagery (Sentinel-1), optical imagery (Sentinel-2, Landsat GLCF, MODIS, etc.), Digital Elevation Models (EUDEM, SRTM, etc.), as well as a series of raster layers such as the Global Human Settlement Layer [7] and the Global Water Surface Layer [6]. In addition, a user can easily create a new collection by importing his/her own data. Any raster collection can be combined with arbitrary vector data sets whether predefined or imported by the user. Examples of predefined vector datasets are: administrative boundaries (GAUL, NUTS, etc.), the Military Grid Reference System used for Sentinel-2 tiling, the Sentinel-2 relative orbits, and the European Natura 2000 protected areas.

3. INTERACTIVE ENVIRONMENT COMPONENTS

The core components of the Jupyter-based interactive environment for geospatial data visualisation and analysis are schematised in Fig. 2. The handling of raster and vector data, processing chains, as well as import/export capabilities are presented in the following three subsections.

3.1. Raster data management

The concept of image collection is inspired by the one proposed on the Google Earth Engine platform [1]. More precisely, the JEODPP interactive library provides an ImageCollection class that allows users to search, select, and filter raster datasets based on a variety of criteria. Users can instantiate an ImageCollection from the relevant dataset (e.g., Sentinel-2 and Sentinel-1) and then choose a geographic location of interest to filter products that intersect a named location (based on calls to the GeoNames online service). Each metadata in

the collection can then be used to refine the selection by using arithmetic, logical, and alphanumeric operators to get the set of images matching all search criteria. For example, all the images acquired in a specific time interval, which have a reduced cloud coverage and have been acquired by a specific sensor on a given relative or absolute orbit can be selected.

3.2. Vector data management

A section of the interactive library is dedicated to the management of vector datasets. Thanks to the mapnik library [5] that allows vector to raster conversion based on rules, vector data are treated in the same way as raster data. The display of vector data can be easily customised by editing all visual attributes (colours, thicknesses, line and fill types, etc.) as well as constructing display legends based on data attributes (single or graduated colour or legends) with colours selected from a vast palette library or directly specified by the user.

3.3. Data processing chains

From the instance of ImageCollection, the user can generate a processing chain by applying data transformation operators to obtain the required analysis and visualisation result. Available operators include the following categories: pixel based operators (e.g., masking, filtering, and band arithmetic), index calculation (NDVI, NDWI, etc.), RGB combination (on-the-fly visualisation of three different processing chains in RGB mode), merging and blending (combination of two or more processing chains using alpha transparency), morphological operators, segmentation, legend management (using predefined legends or creating custom ones from a user-defined list of colours), etc. The resulting processing chain can then be added as a layer to the map and displayed inside the notebook with the ability to zoom and pan. The processing takes place on the basis of the user's display requests: the displayed tiles are calculated in parallel and only at the zoom level required and on the currently displayed extent to achieve on-the-fly rendering even in the presence of extremely complex calculation chains.

More precisely, when adding a processing chain to a map, this processing chain and associated filtered collection are converted to a JSON string. This string is then saved to a database instance linked with a unique identifier (hash code). At the level of the map view, this launches an event to add map tiles based on URLs encapsulating the tile coordinates, zoom level, and a hash code referring to the JSON string defining the required processing. The service responding to this tile request is handled by a Python-enabled web server cluster. The cluster servers read the hash code, retrieve the processing chain definition from the database, apply all processing steps to the selected image data, and compose the map tile that is returned to the IPyleaflet map client where it is displayed. The concurrent map tile requests are already providing some basic parallel processing since multiple requests are triggered in

parallel. In addition, the data reading and processing is performed in a multi-threaded environment where it is possible to make use of the cluster resources to ensure fast responses for the interactive display.

It is possible to build processing chains that integrate raster and vector in the same computing chain. Operations such as masking, selection, filtering, etc. can be applied to combinations of raster and vector data.

All interactive processing and visualisation are performed in the Google Mercator projection at the current zoom level on 256 by 256 pixel tiles. The available input raster data are stored on disk as flat files as downloaded from the respective data source. If needed, faster access can be obtained by converting them in GeoTIFF format with internal tiling and LZW compression. In any case, each single file is complemented by a pyramid representation using overviews as created by the GDAL library. While the visualisation is always based on the production of 256 by 256 pixel tiles, three different schemes are used during processing depending on the type of operations considered:

1. Pixel based operations allow for the 256 by 256 pixel tiles to be processed in parallel and independently;
2. Neighbourhood based operations are addressed by processing tiles in parallel while enlarging them proportionally to the size of the neighbourhood. The processed tiles are clipped accordingly before delivering them the view map;
3. Connectivity based operations such as those resulting from the watershed segmentation or constrained connectivity [10] are handled by processing the whole viewed area and then subsequently tile the results for the view map.

For efficiency reasons, actual image processing is performed through code written in lower level (compiled) languages (C and C++), but this is transparent to the user of the Python package. Functions written in these lower level languages are made available in Python thanks to the automatic wrapping provided by SWIG (Simple Wrapper Interface Generator). This was done for all the functions originating from the *pktools* software suite [4] for processing geospatial data as well as a series of morphological image analysis functions including hierarchical image segmentation based on constrained connectivity [10].

3.4. Data import/export functions

Of great importance to users are the import and export functions of what is displayed and processed. The JEODPP interactive visualisation component can be used to export any processing chain into a georeferenced TIFF image by selecting extent and output zoom level (with some limitations on the total number of pixels involved and a quota system for

storage). This allows users to easily integrate data discovered, analysed, and pre-processed inside JEODPP with external data management and processing solutions, thus allowing better integration and acceptance of the platform. An export method that produces a numpy array out of any band of a processing chain was added, gaining access to a whole suite of powerful data analysis tools. A more evolved product of the JEODPP platform is the ability to export an animation containing a time series. Consider, for example, high-resolution satellite satellites such as Sentinels 2, which, with the recent launch of the Sentinel 2B, can provide an updated image every 5 days (and even less for areas covered by more than one orbit). With only one call to interactive library functions, users can export an animated GIF containing the time sequence of all images on a given geographic location, providing a product of great visual and analytical impact. Also the opposite can be easily achieved: users can upload raster and vector data, in any standard GIS format and SRS, to the notebook management system and get them visualised on the interactive map and combined on-the-fly with other types of data.

4. NOTEBOOK GALLERY

The interactive visualisation and analysis of geospatial data with Jupyter is illustrated in Fig. 3 with three Jupyter notebook snapshots showing the on-the-fly processing and rendering of the European Digital Elevation Model (DEM), the NATURA 2000 vector layers, and the segmentation of Sentinel-2 imagery.

5. CONCLUDING REMARKS AND OUTLOOK

Jupyter offers a very rich environment for interactive visualisation and analysis of raster and vector data sets. The forthcoming operational (v.1.0) *JupyterLab* [2] with its improved interface and user experience will further contribute to knowledge sharing within and across research and governmental organisations. In parallel, the Earth Observation big data shift is calling for the development of protocols and application programming interfaces with other platforms to facilitate cross-platform interactions. In particular, the deployment of some of the functions of the proposed Jupyter based interface on the future Copernicus Data and Information Services (DIAS) will be investigated. Finally, the proposed interactive visualisation and analysis of geospatial data with Jupyter can be used in combination or applied to other types of data. By distributing predefined notebooks it offers an ideal ecosystem for conveying evidence based information in the context of data for policy. Interaction with the data is even accessible to non-programmers thanks to the use of widgets. Extension to other data and application domains for extracting policy relevant information currently include news event and social media monitoring [8] are expected.

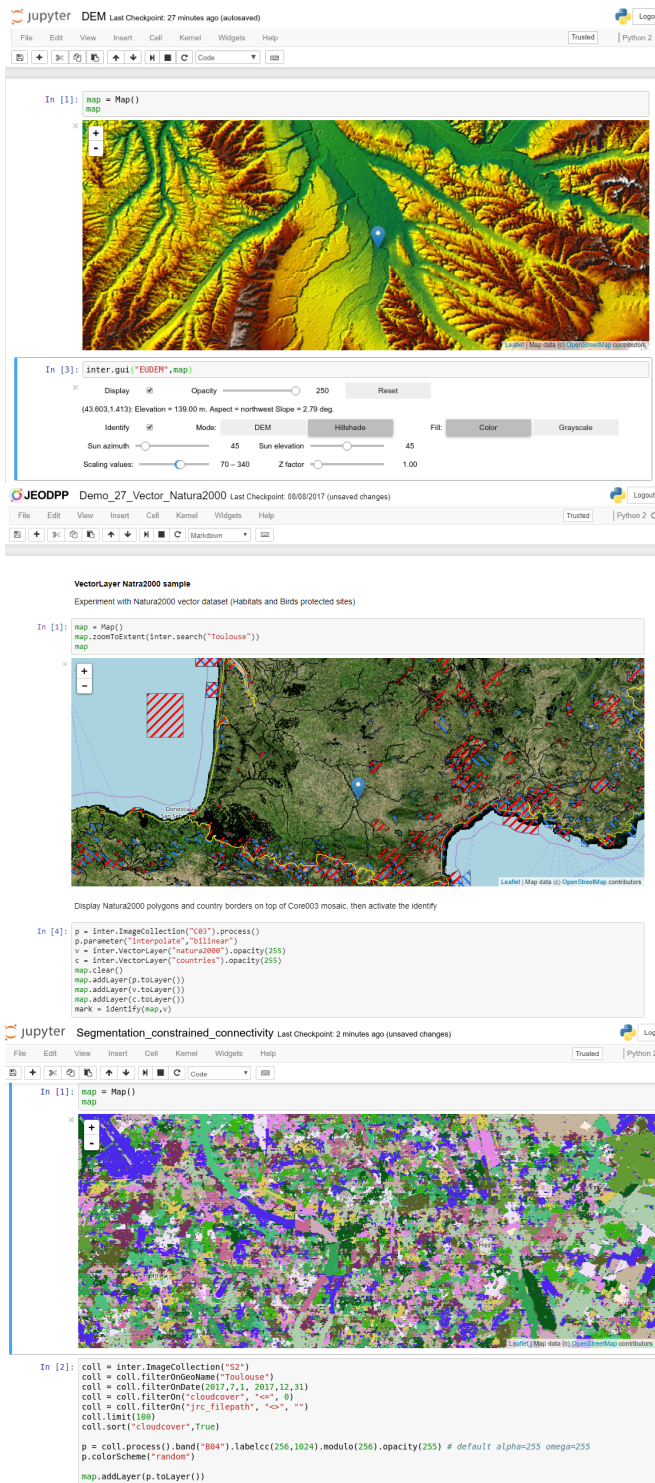


Fig. 3: JEODPP Jupyter notebook gallery with map view over Toulouse. Top: on-the-fly rendering of the European DEM with widgets to control the rendering parameters. Middle: NATURA 2000 vector layer. Bottom: constrained connectivity segmentation of a Sentinel-2 collection on band 4 using a local range set to 256 and a global range set to 1024.

6. REFERENCES

- [1] Gorelick, N. et al. “Google Earth Engine: Planetary-scale geospatial analysis for everyone”. *Remote Sensing of Environment* (2017). DOI: 10.1016/j.rse.2017.06.031.
- [2] Granger, B. and Grout, J. “JupyterLab: Building Blocks for Interactive Computing”. Slides presented at SciPy’2016. URL: <http://archive.ipython.org/media/SciPy2016JupyterLab.pdf>.
- [3] Kluyver, T. et al. “Jupyter Notebooks — A publishing format for reproducible computational workflows”. *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016), p. 87. DOI: 10.3233/978-1-61499-649-1-87.
- [4] McInerney, D. and Kempeneers, P. “Pktools”. In: *Open Source Geospatial Tools*. Earth Systems Data and Models. Springer-Verlag, 2014. Chap. 12, pp. 173–197. DOI: 10.1007/978-3-319-01824-9_12.
- [5] Pavlenko, A. “Open source renders the world”. *Bulletin of the Society of Cartographers* 40.1-2 (2006), pp. 13–16.
- [6] Pekel, J.-F. et al. “High-resolution mapping of global surface water and its long-term changes”. *Nature* 540.7633 (2016), pp. 418–422. DOI: 10.1038/nature20584.
- [7] Pesaresi, M. et al. “Assessment of the Added-Value of Sentinel-2 for Detecting Built-up Areas”. *Remote Sensing* 8.4 (2016), p. 299. DOI: 10.3390/rs8040299.
- [8] Piskorski, J. et al. “Cluster-Centric Approach to News Event Extraction”. In: *Proceedings of the 2008 Conference on New Trends in Multimedia and Network Information Systems*. IOS Press, 2008, pp. 276–290. DOI: 10.3233/978-1-58603-904-2-276.
- [9] Powell, M. et al. “A Scalable Image Processing Framework for gigapixel Mars and other celestial body images”. In: *2010 IEEE Aerospace Conference*. Mar. 2010, pp. 1–11. DOI: 10.1109/AERO.2010.5446706.
- [10] Soille, P. “Constrained connectivity for hierarchical image partitioning and simplification”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.7 (July 2008), pp. 1132–1145. DOI: 10.1109/TPAMI.2007.70817.
- [11] Soille, P. et al. “A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data”. *Future Generation of Computer Systems* (2017). DOI: 10.1016/j.future.2017.11.007.
- [12] Soille, P. et al. “The JRC Earth Observation Data and Processing Platform”. In: *Proc. of the BiDS’17*. 2017.

WEBASSEMBLY FOR EO DATA VALORIZATION, RUNNING (LEGACY) TIME-CONSUMING PROCESSINGS IN THE BROWSER

Nicolas Decoster¹, Julien Gaucher¹, Julien Nosavan²

¹Magellium, Toulouse, France

²CNES, Toulouse, France

ABSTRACT

With the amount of available data from space one needs as much options as possible to bring value to end-users. In that field there is a newcomer: WebAssembly. This technology is a kind of assembly format for running performant processings on any browser. It is a new web standard, with strong support by major web actors, and is a compilation target for existing language (like C/C++) which can brings some legacy code to the browser. This new technology can benefit a lot to online services that manage data: one is no more forced to only execute processings on the server. As a technology which concerns data and processings, it can benefit to Big Data use cases and, in particular, to Earth Observation's ones. This paper presents some experiments that show how WebAssembly can be used for Earth Observation data valorization.

Index Terms— WebAssembly, web standard, browser, processings, online services architecture, JavaScript

1. INTRODUCTION

JavaScript, as a dynamic language, can be slow. Some processings are simply too huge to be run in the browser. And even if processing time would be fine, for most processings there is no implementation in JavaScript, and to be able to use them in the browser one has to rewrite the code. For this reasons, for now, this kind of processings have to be executed server side, which could have an impact on user experience. Well, in fact, until WebAssembly.

WebAssembly (or *wasm*) is a new web standard that is useable now and which is supported by all major browsers' vendors [1][2]. *Wasm* is a low-level binary format that is not meant to be written by hand but is a compilation target. One can see it as a kind of bytecode or assembly language. It lives alongside JavaScript and complements it in terms of processing powers (*wasm* aims near native performance). Moreover *wasm*, as a compilation target, allows the execution in the browser of existing processings written in other languages (C/C++ and Rust for now).

So WebAssembly is a new technology that opens new doors for architecture of online services that manage data, and, in particular, Big Data ones where one always needs to choose the right technologies to properly deliver value to its users. There are lots of scenarios where it can be used. One has limited processing server but its users are ready to host

some processings? One needs to do some complex real-time processing for some interactive data visualization? Some users do not want to upload some of their confidential data on some processing server? One needs a bit more power for a mobile version of a web site or web app? One has an existing image processing algorithm, but is written in C and wants to use it client side? Etc. WebAssembly might help on these cases.

Of course, Earth Observation (EO) and its data with great variety of natures, usages and processings can greatly benefit from *wasm*. This paper first presents what WebAssembly is and where it comes from, and then details some experiments on using *wasm* for EO data valorization with an illustrating proof of concept that integrates image data access (i.e. on the fly decompression), its visualization and some existing or new processings, all in the browser.

2. WEBASSEMBLY

2.1. Origins

Since JavaScript was created, there were lots of attempts to bring processing power or new languages to the browser. There were Java applets, flash animations, various plugins for specific languages (Python or Tcl), integrated new languages (ActiveX, Silverlight, NaCL or Dart), some new standard on existing technologies (WebCL), etc. But all fail to reach major adoption for various reasons: security issues, vendor lock in, poor support by some browser, not well known, not mature enough, etc. Put it in another way: none had a strong enough consensus from the major actors of the web to become the solution for an alternative processing engine to the (sometime) slow and dynamic JavaScript.

But all browser vendors at some time or another had seen that something must done in that field to bring the web forward. In particular, in 2013 some engineers from Mozilla had an idea to bring some power to the web without the need of early adoption from all actors [3]. They observe that what makes JavaScript slow is its dynamic nature which makes it hard for the JIT (Just In Time) JavaScript compiler to optimize execution. They decided to select a subset of the JavaScript syntax which removes all its dynamic nature and to optimize the Firefox JavaScript engine on this subset. They call it *asm.js*. The great idea behind *asm.js* is that as a subset of JavaScript, code written in *asm.js* can execute on any browsers, even those that were not optimize for it. The

other great idea from Mozilla was to work on a tool chain, emscripten, which compiles C/C++ code either to JavaScript or to asm.js (actually emscripten was created before asm.js). At the time it was a kind of foolish idea, but when they test it, it appears to work and perform very well, with some processings that even run at near native speed.

Asm.js was the first technology that combines all the required features: it works everywhere, it is a compilation target and it can have near native performance. It was clear that it was a serious candidate to bring processing power to browsers. All major browsers vendors saw that and in 2015 decided to move forward in that direction together. The idea and the dynamic behind WebAssembly have started [4].

2.2. Description

WebAssembly is a web standard which first version is implemented and usable now on the four major browsers (Firefox, Chrome, Edge and Safari). It is a kind of assembly or bytecode. The standard defines a binary format, a text format and a JavaScript API to load, compile, instantiate, manage and run wasm modules from regular JavaScript code and to add interaction between the JavaScript engine and the wasm's one. Wasm is secure by design and has the same level of security than JavaScript.

One usually does not write wasm code directly but use tools to compile code from a given language into binary wasm code. For now, efforts are concentrated mainly on providing a good tooling to compile C/C++ using a SDK called emscripten which is based on the Clang compiler and the intermediate representation LLVM. Emscripten provides very useful utilities like browser implementation of some C features (like file I/O on in-browser virtual file systems) or a fallback to WebGL for OpenGL code. The SDK provides tools to execute build steps (like configure and make) of existing libraries in the context of emscripten builds.

2.3. Usage

WebAssembly is a new technology in the tool belt of architects that build online services that manage data. Before it, if there was a need for some time consuming processings the only option was to run them server side. But sometime running server side is not an option, for interactive reasons or simply because one cannot afford some server side processing power or because of poor network communications. Wasm can help on these cases.

With wasm one can use some legacy code in the browser as long as it is written in one of the languages that compile to wasm (C/C++ or Rust for now). Of course, there are some constraints on that code and its dependencies (one cannot compile code that calls low level features, like hardware communications) but the current tooling does a good job and relatively complex libraries can actually be compiled and executed browser-side.

Let us mention two use cases for wasm in Big Data context. First, wasm can be used to execute some processing

on some data extract in the browser, letting a user tries some different parametrizations, visualizes the results, tries some more runs and, when the results are fine for her, launches exactly the same processing with the parametrization she chooses on a much bigger dataset on the server. Another scenario is to use the browser to mashup some data from different sources, using some very special processings that are not offered by the corresponding servers, and with some huge local files in specific format that the browser might not understand by default (like JPEG 2000).

3. EXPERIMENTS: USE OF WEBASSEMBLY FOR EO IMAGE VISUALISATION AND PROCESSING

3.1. Presentation

This section presents some experiments made with wasm for EO image visualization and processing in the browser. The foundation of these experiments is a proof of concept of a web app where everything is run in the browser. This tool uses OpenLayers for the visualization part and modern web technologies are used when appropriate (web workers, ECMAScript 2015, IndexedDB, File API...).

The use case is: how one can view images in format not recognized by browsers and run on them some custom or legacy C/C++ processings? This work is focused on JPEG 2000 (JP2) format and on plane and cloud detection using simple classic algorithms or machine learning inference.

The challenges are the compilation of existing (possibly complex) libraries into wasm and their seamless integration with non-wasm parts into the web app while keeping a good user experience on a standard laptop.

3.2. Viewing JPEG 2000 images in the browser

The objective of this part is to allow a user to drag a JP2 image from its computer into the web app and to navigate into it smoothly as any other OpenLayers layer (say, like any WMTS layer). To achieve this, one needs to efficiently bring some file content into the browser, use some JP2 decompression algorithm and integrate it with OpenLayers.

First, we investigated two libraries for the JP2 decompression part: OpenJPEG and Kakadu. Both have been successfully compiled into wasm with emscripten after some tweaking of the building process. This compilation step brings the decompression executables into the browser (`opj_decompress` for OpenJPEG and `kdu_expand` for Kakadu). Some informal test showed that Kakadu is faster, so we chose it for integration into our app.

We then needed to bring the JP2 image content into the browser. For that, classical web API were used for the drop capture and for the creation of a File object into the browser.

At last, for the integration of `kdu_expand` with OpenLayers, we have created a custom layer that calls the decompression for each tile that the view requests for the current viewport. So, before each time a new tile is

displayed, a `kdu_expand` command is executed in the browser to read and decompress the JP2 content.

Tests with a 30 MB mono band Sentinel 2 JP2 file show that the navigation is as smooth as with a Bing layer that requests tiles from the network. Note that `kdu_expand` accesses file content using an emscripten pseudo file system which resides in memory (called MEMFS). This means that the entire file must be loaded in the browser's memory. That's fine for relatively small files (we succeeded with 130 MB files) but lead to a browser's allocation error with bigger ones (1 GB in our case). Luckily, emscripten provides a second kind of pseudo file system (WORKERFS) which provides read-only access to File objects inside a worker without copying the entire data into memory. So we included our `kdu_expand` into a web worker and put our big JP2 into this pseudo file system. There are no more allocation errors and the big JP2 file is displayed correctly even if not as smooth as the one from pure memory.

These results are very promising. The display from memory is already usable and there is room for amelioration as only the default `kdu_expand` implementation was used. A custom use of Kakadu, best suit to OpenLayers needs, might bring some performance improvements.

3.3. Plane detection

As a first experiment of using wasm for legacy processings in the browser, this work has investigated the detection of plane in EO images. This use case is based on the ICF (Integral Channel Features) machine learning algorithm [5] from the CCV library. ICF is one of the algorithms used in an internal work by Magellium and CNES to study how some computer vision algorithms for object detection can be used on Pleiades images [6]. We choose ICF for our wasm experiments because it showed some good results and CCV library is developed in C with little dependencies. We used the ICF plane model trained during this internal study.

First, the inference executable of ICF algorithm has been compiled to wasm using emscripten with some minor tweaking of the build process. The integration with OpenLayers consisted of getting the image content of the canvas of the map view, storing it on the emscripten pseudo file system, executing the ICF inference on this content, parsing the detection results and displaying them as circles on a new layer. Each detection comes with a score, and the circle color is set depending on whether the score is lesser (negative detections) or greater (positive ones) than a threshold. One can imagine that a user executes the ICF detection on the current view, then reviews each detection and finally use a mouse click to switch positive/negative status based on visual inspection. Our web app has these features implemented to illustrate this use case.

The inference in the browser executes in one to ten seconds on a typical view. Informal tests show that some native execution (from an x86 executable) of ICF inference can be two times faster. By default wasm execution is done

in the main thread of the JavaScript engine. So during its execution the browser is freeze, which impacts dramatically the user experience for processings that take more than few milliseconds (like ours). To address this issue one usually uses web workers. So, we embedded the inference into a worker, and switch to another emscripten file system (IDBFS) which uses IndexedDB as a storage backend that can be shared between the main thread and any worker.

3.4. Deep learning inference

Deep learning for EO is gaining more and more interest. And one might need to execute some inference in the browser. Some EO use cases are object detection (same use case as plane detection above) or for image segmentation like cloud labelling or land use classification. There are progresses on this subject every day and Magellium actively works on deep learning technics, and, in particular, on segmentation. While this is a work in progress, during this work we have investigated the use of wasm to execute deep learning inference. So far, we have used WebDNN library to successfully compile one of our neural network for segmentation into wasm and execute it in the browser on a Sentinel 224x224 tile but using some random network's weight. We had to adjust the network to WebDNN which, for now, missed some types of network layers. These marginal changes explain why useable weights are not available yet, as we cannot use the weights of our successful trainings on this slightly different network. As the time of this writing, there is some ongoing work to train this network on some use cases like the ones listed above. So, even if actual segmentation results cannot be shown now, we are confident that we will soon successfully use wasm deep learning inference to produce some segmentation map in the browser. Our successful experiments with WebDNN show that this is technically possible.

3.5. WebAssembly on the server

WebAssembly was designed mainly with the browser in mind. But it is not specifically bound to some browser exclusive technology. In fact, the only integration in the browser is with the JavaScript engine (but even this integration is not mandatory). And Node.js, as a JavaScript environment based on a browser JavaScript engine, naturally embeds a wasm engine too (starting at version 8). So, with Node.js, it is also possible to use wasm code on the server, using exactly the same integration than in the browser. This way, one can develop some processing, compile it in wasm and use it exactly the same way on the client and server sides. This eases deployment and guarantees that exactly the same binary is used on both sides. As a bonus, one gain the security level of the wasm engine for the processing on the server which can act as a kind of sandbox.

This work has investigated the use of WebAssembly on the server by extending the plane detection use case: a user

opens a big image in the browser and does some experiments on a part of the image with the ICF plane detection (like testing some pre-processing or using different resolutions). Once she is satisfied with the result, she executes the detection on the full image and maybe on some other big images. Of course, for this step it can be better to process server side, maybe on some kind of cluster.

For this, one can implement a simple web service where any user can post some processing request for the plane detection. We have experimented this by using a serverless architecture. We chose Google Cloud Functions (GCF) but any other serverless solution will do. We created a GCF that embeds our wasm plane detection and that is triggered by a HTTP request. With this, one can send any number of requests to execute the processing in parallel on a bigger data set. And exactly the same binary is executed during the tuning and the batch processing.

3.6. Performances

Even if WebAssembly “aims to execute at native speed” [1], it can already be very useful if it executes significantly faster than JavaScript. The idea is that wasm should bring into the browser some use cases that were not possible in plain JavaScript. And, as a bonus, it would be nice to have near native speed. This way, processings that were bound to target native execution because of some performance requirements could then target wasm, and then can be executed in the browser.

Actually, for now, to our knowledge, there is no complete study that benchmarks wasm against JavaScript and native. But every now and then people conduct some informal speed tests and they tend to confirm the good performance of wasm, which can be significantly better than JavaScript and often has near native speed. Our own investigations confirm this.

We have implemented some basic algorithms based on array manipulation (the most complex one is an image convolution) in C and in JavaScript. We have compiled the C code with emscripten to produce wasm and JavaScript (this JavaScript code is not the same as the JavaScript code written by hand above, it is an output of the emscripten compiler). We have also compiled the C source code with GCC and Clang. The benchmark shows that wasm execution is always faster than the JavaScript implementation’s one and sometimes significantly faster (more than 20 times faster for some convolutions); wasm is actually near native speed (in fact differences between wasm and Clang are in the same order than the differences between Clang and GCC) and sometime it is faster (especially on convolutions). More surprisingly, the JavaScript produced by emscripten shows very good performance too and can be faster than wasm on some cases, and even than native on other cases.

As always with benchmark, one has to be careful when extrapolating to her own use case and has to conduct her

own benchmark before drawing conclusions. Note also that some other aspects have impact on performances: code size and transfer, code compilation by the browser (either JavaScript or wasm), communication between JavaScript and wasm engines, etc. But we can say that the web platform as a whole actually has now very good performances and that wasm is a serious option in terms of performance. Note also that wasm is a young technology in terms of standard and implementations (the first official browsers’ engines landed in March 2017) and one can expect better and better performances in the future.

3.7. Other experiments

During our experiments we also successfully compile GDAL with emscripten and use it in the browser. We developed and compiled in wasm a cloud detection algorithm which successfully runs interactively on the viewport each time the user changes a threshold with a slider. We have done some integration work of emscripten in the browser like developing a tree-view of the emscripten file systems and a shell-like command line tool in the browser to navigate through emscripten file systems and to execute wasm executables (like `kdu_expand`).

4. CONCLUSION

In this paper some experiments have been detailed that show the potential of WebAssembly and emscripten for EO data valorization. They bring one’s tools in the browser and make them fast, and are worth considering as complementary technologies that open new doors for the design of web based data intensive systems.

5. REFERENCES

- [1] ‘WebAssembly’. <http://webassembly.org/>
- [2] A. Haas et al., ‘Bringing the Web Up to Speed with WebAssembly’, in Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation, New York, NY, USA, 2017, pp. 185–200.
- [3] ‘asm.js in Firefox Nightly | Luke Wagner’s Blog’. <https://blog.mozilla.org/luke/2013/03/21/asm-js-in-firefox-nightly/>
- [4] ‘WebAssembly | Luke Wagner’s Blog’. <https://blog.mozilla.org/luke/2015/06/17/webassembly/>
- [5] P. Dollár, Z. Tu, P. Perona, and S. Belongie, ‘Integral Channel Features’, in Proceedings of the British Machine Vision Conference, London: BMVC Press, 2009, p. 91.1-91.11.
- [6] Nicolas Decoster, Grégory Loeb, and Julien Michel, ‘R&T CNES détection d’objets - Rapport d’étude’, private communication, 2016.

INTEGRATION OF WEB WORLD WIND AND SENTINEL HUB - A GLOBAL 4D BIG DATA EXPLORATION AND COLLABORATION PLATFORM

Grega Milcinski¹, Guenther Landgraf², Patrick Hogan³, Paulo Sacramento⁴

(1) Sinergise, Ljubljana, Slovenia

(2) ESA, European Space Agency, Frascati, Italy

(3) NASA

(4) Solenix Deutschland GmbH, Spreestrasse 3, D-64295 Darmstadt, Germany

ABSTRACT

The volume of open Earth observation data, available to the world, has grown significantly over the past few years. Copernicus Sentinel satellites joined Landsat and other Earth observing missions, and now Copernicus has become the largest single Earth observation programme in the world, as well as the third largest data provider. The field of Earth Observation (EO) has changed dramatically and done so almost overnight. Previously the major challenge was "where to get remote sensing data?" Now it is more about "how to make use of this vast abundance of available data?" To make use of this data, most of the processes in the past were manual or semi-manual. Now a statistical approach is used for how to most effectively process all of the data, for interpretation via visualization in an efficient and intuitive way. To achieve this, we have integrated Sentinel Hub, a Copernicus award-winning cloud-based satellite imagery processing and distribution service, with WebWorldWind, a 3D virtual globe visualization environment. Given the temporal aspect to EO monitoring, we provide a 4D big data exploration and collaboration web app, where the user is able to analyze multi-spectral satellite datasets together with other spatial datasets.

1. SENTINEL HUB

Sentinel Hub [1] is a satellite imagery processing and distribution service. Its technology is optimized for on-demand real-time processing of satellite's big datasets. This makes it possible to effectively and efficiently leverage open data distributed by several cloud providers - instead of the standard pre-processing steps, performed by typical platforms (e.g., re-projection, tiling, building true-color composites, etc.), this platform does practically all the steps on-the-fly, as the user requests the data. There is only one pre-processing step, the indexing of data and their associated files, making access to the data much faster. There are two important benefits of this approach. First is cost-of-operation, which is an order of magnitude less than similar systems, due to the fact that costs occur almost solely when the user requires data specific to their area of interest. Our unique approach solves a common problem with EO data, namely that users are only interested in small percentage of data available, but it is impossible to

accurately predict, which ones are those areas of interest. The second benefit, possibly even more important, is flexibility of the service. By avoiding time-consuming dataset processing tasks, it is possible to add new features to the platform in a matter of minutes. It also gives the user flexibility for what kind of actions to do with the data.

1.1. Basic features

The core feature of Sentinel Hub is to bring big EO data to the data-user's environment, one request at a time, making it easier to integrate this data into existing applications, using supported OGC standard web-services (WMS, WCS, WMTS). Such interfaces are well-known in the geographical information service community and are an important step to "bring space out of space", making EO space data easily accessible to the non-space industry. For advanced tasks there are dedicated APIs used to get a statistical analysis of a specified area over time. Users can choose among standard composite configurations (true color, false color, NDVI, EVI, NDWI, etc.) but, more importantly, they can create their own combinations using either simple band-mapping or even band-math. By using the power of ongoing monitoring of Sentinel, Landsat and similar missions, it is possible to construct a viewport that stitches together scenes over a specified time interval. This way, users have full control across processing steps, projection type, methods for down- and up-scaling, order of mosaicking, and so much more.

1.2. Supported satellite missions

The prime focus of Sentinel Hub are optical missions, currently supporting Sentinel-2, Sentinel-3 (OLCI), Landsat-5, 7 and 8, Envisat MERIS, RapidEye, Planet Doves, Pleiades and the MODIS Terra and Aqua data. Complete archives of open-data missions can be previewed within the EO Browser [2]. Access statistics and comparison with more isolated approaches experimented with using ESA heritage missions have shown that use of historical data series get significantly incentivized by a service co-hosted with compatible fresh data, namely from the Sentinels.

The level of effort required to support another data source is quite manageable, providing that data are available in one of

the clouds. The steps required are the analyzing of source files (GeoTiff, JP2, NetCDF, etc.) to identify relevant chunks, meta-data files (acquisition information, quality attributes, etc.) and overall satellite constellation parameters (spectral bands, revisit-times and related criteria).

Recently the platform was upgraded to support SAR data as well, starting with Sentinel-1, currently in beta operation. Aside from the new data formats, a challenge needing to be solved was on-the-fly orthorectification. Standard Sentinel-1 distribution comes with non-orthorectified data, given the high cost of processing all scenes. So, a run-time optimized process was implemented, using data fusion with DEM data, residing on the AWS platform.

1.3. Advanced algorithms and temporal analysis

Being able to stream data to the user's environment is a first step, but not necessarily a sufficient one. For efficiency and performance, users need to process several steps on the platform itself.

Ad-hoc algorithm support is an extension of standard-band math. By using JavaScript as the main interface, users can script their own steps, e.g. implementing Hollstein's cloud detection method [3] in a matter of minutes.

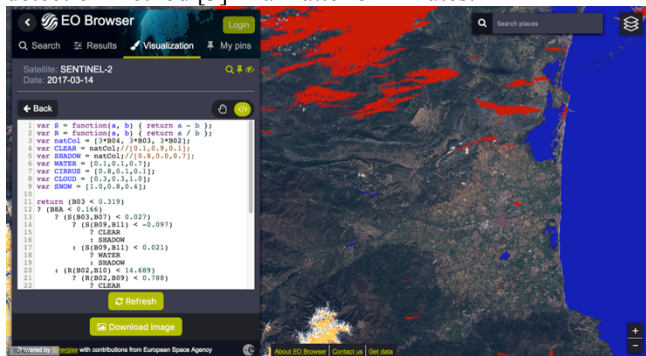


FIGURE 1 - EO BROWSER RENDERING DATA WITH HOLLSTEIN'S CLOUD DETECTION ALGORITHM

The power of custom scripting becomes apparent when temporal component comes into focus. In addition to scene-based mosaicking, by using the meta-data, one can work on a per pixel basis, analyzing all available information in a chosen time range. Simplest result is selecting best pixel, for cloudless mosaic, or maximum NDVI over one quarter, to cluster data in specific area. One can stretch this further, e.g. observing index variance over time, comparing this to typical temporal signatures, essentially classifying land cover. Or searching for drop in index value to identify peak time for harvesting or burned areas.

2. ARCHITECTURE

2.1. Cloud environment

The most important input for Sentinel Hub is having fast access to the un-packed source satellite data, with sufficient speed and capability to read parts of the files, thereby

contributing significantly to performance. The first working environment was AWS cloud, with an easy and open access to various datasets [4]. After establishment of EO Cloud Platform [5] Sentinel Hub was able to access even more ESA archives and provide a long optical time series, including Sentinels, Landsat and Envisat MERIS. In the near future, another important platform should be available - DIAS - which portends even more relevant data being accessible for Sentinel Hub.

2.2. Data structure optimized for cloud processing

EO data processing is changing as we have data readily available "nearby". The concept of "scene" is no longer relevant. People often process only small parts of it (e.g. amazing 10m resolution of Sentinel-2 does not help much when observing 10.000 sq. km at a time) or several scenes at a time (at scene borders). A temporal component is becoming more and more important. It is therefore essential to be able to access relevant part of the data as fast as possible. Cloud optimized GeoTiff [11] is an initiative trying to standardize the main concepts, to ensure fast processing without adding significant volume or requiring intense computing resources. The essential parts are internal tiling (so that pixels related to nearby area are packed together) and file content information (header) at the beginning of the file, so that one can immediately find relevant parts. Overview layers in the file are welcome, to speed up low-resolution preview. The most important part however is that files are available in an unzipped form in object storage. Sentinel-1 files produced by ESA are in GeoTiff, with the header part at the end of the file, so there is no internal tiling nor overview layers. To make these data useful on AWS, we had to re-encode them. Perhaps of interest, is that after using lossless compression, the resulting files are about 15% smaller than originals even though they contain overview layers. JPEG2000 used by Sentinel-2 is better organized and can be used as is, although up-front indexing is required and internal tiles could be larger. Sen2Cor processed S-2 is unfortunately a step back. These are important points, which should be considered by future data providers as well as DIAS operators. If these fail to ensure fast access capabilities, platforms will not be interesting for large scale operations.

2.3. One second from a request to the result

Standard operation requires execution of several consecutive steps. Based on the request parameters (area, time range, cloud coverage, etc.) we select relevant scenes based on meta-data stored in PostgreSQL. These are then combined together based on "priority" parameter, e.g. most recent on top. The viewport is therefore filled with scenes, each part with exactly one source, with slight overlap occurring only on internal borders to prevent gaps. The next step is getting relevant data for each of the bands used in processing. This process is extremely optimized, downloading only necessary

parts. After decompression we apply pre-mosaic filters (on-the-fly atmospheric correction, radiance to reflectance operations, etc.), combine the scenes using user-configured math band, add post-mosaic filters when needed (dynamic contrast, HDR and similar), re-project the result in CRS of choice and encode it in user-defined output format (GeoTiff, JP2, PNG, KMZ, etc.).

Orchestration of these steps is extremely important to ensure fast operation. Practically all parts of the process have been written from scratch for performance (decoding being an exception). Optimization requires adaptation of the process for each new data source. However, as there are limited variants of these compared to the unlimited amount of data, this effort makes sense. It becomes even more important in multi-temporal processing, as it is common to process couple of tens of scenes before one can produce a meaningful result. And we can parallelize this process to make it an order of magnitude faster.



FIGURE 2 - TURKEY, GREECE AND CYPRUS CLOUDLESS MOSAIC PROCESSED BY PIERRE MARKUSE USING SENTINEL HUB MULTI-TEMPORAL PROCESSING

To achieve fastest access of the data, we operate several environments, one at AWS EU, one at AWS US-West, one at EO Cloud in Poland. This is opening the door to data fusion. The first example was orthorectification of Sentinel-1 data, which is happening on Polish EO Cloud and uses the digital elevation data from AWS US in real-time.

2.4. Economy of operation

An important factor of Sentinel Hub's success are its low costs of operation. As it works with source data, costs of storage is insignificant, e.g. there is no tiling and pyramids, neither in pre-processing stage nor during operation. File indices for Sentinel-2 take less than 0.1% of the volume. We do generate orbit previews for extremely low scale analysis (e.g. continent scale, where one would need to process hundreds of scenes, too much for real-time operation) but these also take less than 0.3% of the volume. Software optimization made it possible to run services on standard VMs, one being able to process several concurrent requests by itself. The entire operation is therefore extremely cost

efficient, which obviously depends significantly on a freely accessible data archive.

3. WEB WORLD WIND

Before there was a Google Earth, there was WorldWind. The world's very first open source virtual globe platform is today's ESA-NASA Web WorldWind (WWW). WWW is the web version of WorldWind, started in 2014. It is a free and open-source 3D virtual globe API for HTML5 and JavaScript. It is based on WebGL, available stably as part of HTML5 since 2011 (also on mobile platforms). Being web technology, it is great for distribution, portability, data access and ease of use. ESA joined the NASA WWW effort in 2015, bringing with it development capabilities and requirements.

This browser-based 3D Software Development Kit (SDK) is now, therefore, being jointly developed by NASA and the European Space Agency. The GIS community could not have two more relevant government agencies advancing the ideal virtual globe platform for realizing spatial data. This is open source at its very best, an easy-to-set-up and easy-to-use tool, which can be used to power any number of spatial data management applications. WWW is used extensively by U.S. Government agencies, including the Federal Aviation Administration, and now the European Space Agency, including in the popular ESA Sentinel App since the Summer of 2016 [9].

The ESA and NASA development teams are jointly advancing the WWW virtual globe platform with one mission in mind, a place for the world to work with Earth observation data [7]. Next features being developed now include incorporating the npm package manager for the JavaScript runtime environment Node.js, and the ability to experience time series data via WMTS. Additionally, an easy-to-install WMS server using GeoServer, that includes the base set of Earth data for imagery and terrain is almost complete [8].

The ESA-NASA collaboration has enabled a major boost to the feature set of WWW, which now includes:

- Support of standard formats: JPEG, PNG, GeoTIFF, Shapefile, GeoJSON, KML (geometries, placemark, styling, overlays, time primitives, network links), Collada (for 3D models)
- Support of OGC standards: WMS, WCS, WMTS, WFS
- Aesthetic features: Atmosphere, Day/Night effect, Skybox, Antialiasing
- Utilities: Layer Management, Controllers, Recognizers
- Others: Measurement, Analytic Surface, HDF, FBX, Starfield (also Sun, planets), Performance improvements & optimizations, Camera model/controller, Shape Editor, WKT



FIGURE 3 – WEBWORLDWIND IN ACTION

4. INTEGRATED PLATFORM

OGC WMTS was chosen for integration due to WWW's strong support for open standards. The process is straightforward - one gets an account for Sentinel Hub, configures desired layers of the WMTS instance, and uses the provided WMTS URL within WWW. The GetCapabilities OGC function is used to recognize available layers, which are automatically offered to the WWW user. It is also possible to control the temporal aspect of the EO data directly within WWW by choosing the desired date. A demo app illustrating this process and showcasing the potential of the Sentinel Hub, particularly for what concerns Sentinel data, is now included as part of WebWorldWind's main code base [10], being therefore available to any developer who gets and uses the framework. This is made possible through the usage of a demo key provided free-of charge by Sinergise, other keys can be used for building actual applications.

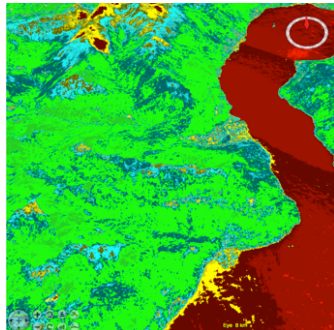


FIGURE 4 - LAGO DI GARDA, VISUALISED WITH MOST RECENT NDVI DATA [6]

5. PLATFORM FOR ANY USE-CASE

The github website where the suite of WorldWind platforms calls home, provides virtual globe solutions in Android and Java, as well as the Web version. Google Earth is just one application, WebWorldWind provides the GIS community with any number of Google Earth like applications. And because of the Sentinel Hub's affinity for Web WorldWind, it is straightforward to access Sentinel Hub web services and instantly have global coverage.

6. FUTURE STEPS

Both platforms - Sentinel Hub and WWW - are rapidly developing as stand-alone solutions, and more recently in a

highly orchestrated manner. The Sentinel Hub is focusing efforts on advancing temporal processing, both from performance point of view as well as adding new functions, e.g. efficient scene selection, etc. A major step forward currently being explored is integrating supervised and unsupervised machine learning.

ESA and NASA are currently working on advancing a variety of new features of WWW, with some only in the planning stage. These include, amongst others:

- OpenSearch for EO
- Enhancement for high-density screens, e.g. Retina
- Vector graphics for icons, e.g. SVG
- Loading performance, e.g. base 64 encoding for small images and image sprites
- Heatmap/Density Layer
- WCS Elevation Model (without hardcoded server)
- Web Processing Service (WPS)
- Line-of-Sight
- GeoPackage
- Deep Picking
- Surface Shape
- Surface Placemark
- Surface Text
- KML Expansion & Refinement
- Shape Editors
- Path Editing
- Polygon Editing
- API Documentation
- Instrumentation
- Unit Tests
- Shape Labels
- SVG Image Source
- base64 Image Source

7. CONCLUSION

Vast amount of remote sensing data acquired on a daily basis does not benefit anyone unless that data is accessible and usable – meaning readily analyzed, visualized, easily shared and interpreted by the respective user communities. Sentinel Hub, an efficient processing tool, powered by machine learning capabilities, when integrated with the open-source 3D/4D visualization technology WebWorldWind, becomes a perfect platform for research and commercial activities. The fact that these services and technologies are strongly supported by ESA and NASA as well as the open-source community, assures their vibrant development and long-term sustainability.

8. REFERENCES

- [1] <http://www.sentinel-hub.com>
- [2] <http://apps.sentinel-hub.com/eo-browser/>
- [3] <http://www.sentinel-hub.com/blog/ad-hoc-testing-algorithms-globally>
- [4] <https://aws.amazon.com/earth/>
- [5] <http://www.cloudferro.com/en/eocloud/>
- [6] http://thales-geo.github.io/webworldwind-demos/WWD/apps/WMTS/WMTS_ESA.html
- [7] <https://github.com/NASAWorldWind>
- [8] <https://github.com/NASAWorldWind/WorldWindServerKit>
- [9] <https://play.google.com/store/apps/details?id=esa.sentinel>
- [10] <http://get.solenix.ch/wmts-demo/apps/SentinelWMTS.html>
- [11] <http://www.cogeo.org/>

OGC BIG DATA WHITE PAPER- ABSTRACT

Marie-Françoise Voidrot, George Percivall

Open Geospatial Consortium

ABSTRACT

The Open Geospatial Consortium (OGC) Big Geospatial Data White Paper is a survey with these main themes:

- Geospatial data is increasing in volume and variety.
- New Big Data computing techniques are being applied to geospatial data.
- Geospatial Big Data techniques benefit many applications.
- Open standards are needed for interoperability, efficiency, innovation and cost effectiveness.

The main purpose of this White Paper is to identify activities to be undertaken in OGC Programs that advance the Big Data capabilities as applied to geospatial information.

This white paper was developed based on two Location Powers events organized by OGC in 2016 and 2017.

Index Terms— Big Data, OGC, White Paper, Open Geospatial Consortium, OGC documents, bigdata, geospatial, location, standards

PREFACE

This paper summarizes the “Open Geospatial Consortium (OGC) Big Geospatial Data White Paper” [1]. This white paper was developed based on the two following Location Powers events [2]:

- Location Powers: Big Data, Orlando, Sept. 20th, 2016 [3].
- Location Powers: Big Linked Data, Delft, March 22nd, 2017 [4].

1. INTRODUCTION

Every two days the human race is now generating as much data as was generated from the dawn of humanity through the year 2003 [5]. Out of its continuing task of analyzing the future technological trends, the Open Geospatial Consortium has identified a use and needs for open standards to support interoperability, efficiency, innovation and cost effectiveness to better leverage the use of new big data computing technologies to geospatial data to benefit many applications.

Via two main « location powers » events, the OGC brought together leading developers of big geospatial data

systems to advance collectively the use of big data computing techniques applied to geospatial data with results in many applications.

Based on these works, the OGC Big Data white paper:

- describes the value of Big Data techniques to several end-user applications, and
- highlights the commonalities between Big Data use cases to reduce the complexity in applying the technology, and
- presents several high priority focus areas for advancing big geo data implementations based on open standards as opportunities for OGC activities, and
- lists existing and potential new activities for consideration to be undertaken in OGC Programs and in coordination with external alliances.

2. THE VALUE AND USE CASES OF BIG GEO DATA APPLICATIONS

New technological means allow to generate and collect a huge amount of data. Among them geospatial data have a unique place and their impact is related to their distance to the point of interest.

Earth Observations, Resource Management, Mobile Location Services, Transportation and Moving objects and Smart Cities provide good examples of geospatial applications using Big Data techniques. These applications require “extensive datasets — primarily in the characteristics of volume, variety, velocity, and/or variability — that themselves require a scalable technology for efficient storage, manipulation, management, and analysis.” [6].

Adding a “Modeling and Simulation” group to a reference architecture under development by ISO/IEC JTC 1/WG 9 – Information technology – Big Data [7], the paper shows the value of these new capabilities for such applications by enhancing the high commonality of several use cases. These use cases are organized into four groups as shown on Figure 1:

- “Collection and ingest” of Big Data requires the abilities of high velocity data streaming.
- “Data preparation and structuring” involves processes that convert raw data into cleansed, organized information increasing the structure of source data in a pre-defined data model or in a pre-defined way.

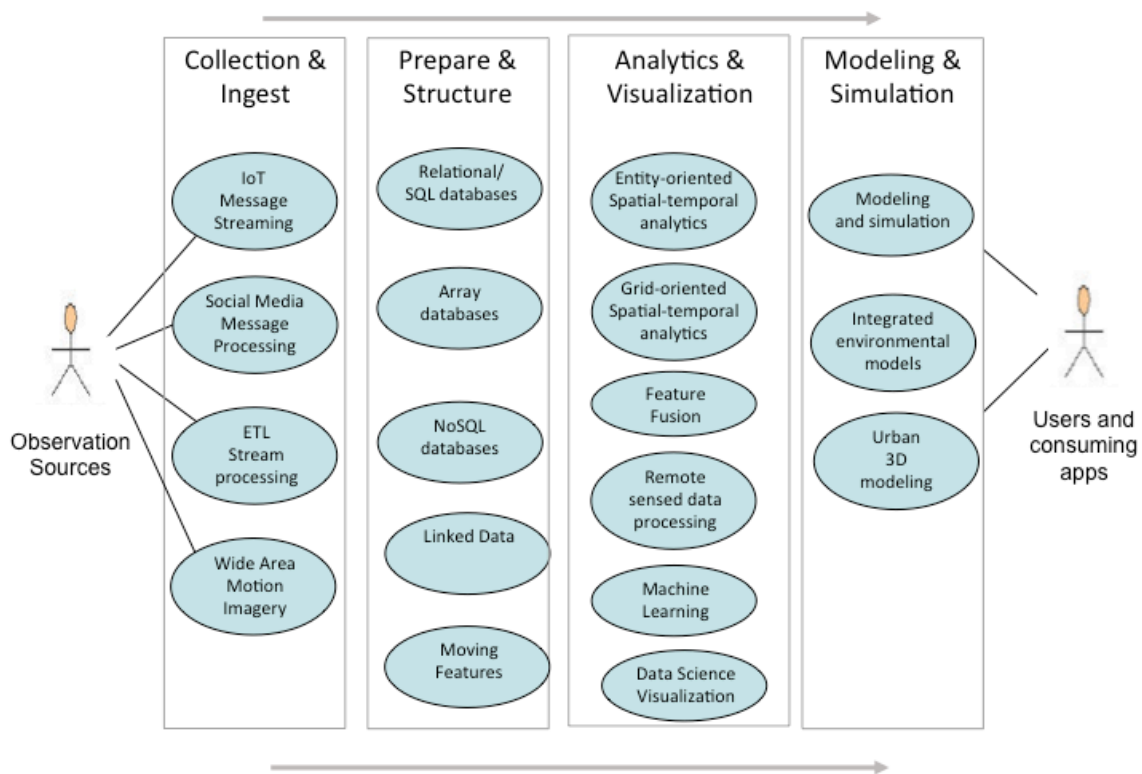


Figure 1: Big Geo Data Use Cases

- “Analytics and Visualization” implements techniques to extract knowledge from the data based on requirements that specify data processing algorithms to produce new insights that will address technical goal of a vertical application.
- “Modeling and Simulation” focuses on insights into the interaction of the parts of a system, and the system as a whole, to advance understanding in science and engineering and inform policy and economic decision-making.

3. OGC BIG GEO DATA OPPORTUNITIES

OGC by its mission, supports the development of geospatial processing based on open standards. The need for open standards in Big Data has been recognized in the wider Big Data developments. For example, “Use of standards and related issues in predictive analytics”, a presentation by Paco Nathan, O’Reilly Media at the KDD conference, 2016-08-16 [8] identified a Lesson from the success of Apache Spark is “lack of interchange for analytics represents a serious technical debt and potential liability”. Based on the discussions of the two Location Powers events as well as in the OGC Big Data Domain Working Group, the paper lists several implementations and ongoing works on Cloud computing for EO data:

- in OGC Testbed 13 framework to address the major issues of Cloud API interoperability and

application portability from one cloud to another, (OGC Testbed-13 supports the development of ESA’s Thematic Exploitation Platforms (TEP)) (See Figure 2).

- an approach from DigitalGlobe of pre-computing and storing image chips in an object store to supports analysis ready paradigm.
- the SciSpark project by JPL [9] that marries Apache Spark with climate science into a scalable system.
- a commercial project of hosting of Exelis ENVI Analytics on DigitalGlobe’s Geospatial Big Data Platform to extract information out of a 15-year catalog of high-resolution satellite imagery.

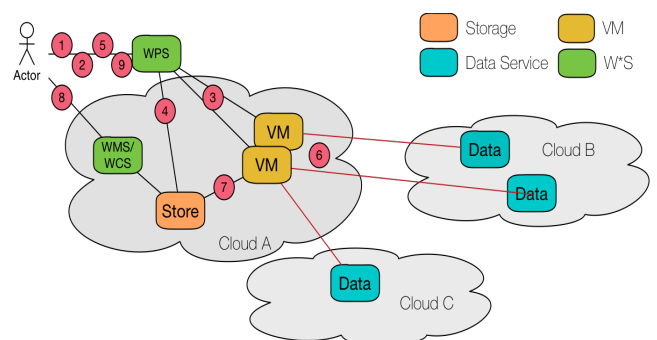


Figure 2: OGC Testbed 13 Cloud Environment Overview

The OGC Big Geo Data white paper also highlights the themes of “Analysis Ready Data (ARD)” and the current implementations of Data Cubes: the Australian Geoscience Data Cube [10], the Earthserver project [11], the CEOS Open Data Cube (ODC) [12] and the QB4ST [13]. Works are going to progress towards Data Cubes interoperability.

Placing the existing data representation methods for geospatial information into a big data context, OGC defines a base for its big data activities on several data representations: Features, Coverages and DGGS as a new approach of Coordinate Reference System (CRS). Ideas like bringing OGC Simple Features to Big Data world’s “Data Frame” object types are being considered. Recently, the standards and implementations of Web Coverage Service (WCS), Coverage Implementation Schema (CIS) and the Earth Observation profile of WCS have been extended to 3D and 4D and coverage collections have been defined to handle a large number of layers providing the consumers facilities to request in 3D/4D domains and receive N-Dimensional range/feature data. Last but not least, OGC has recently approved the OGC Discrete Global Grid System (DGGS) Core Standard [15-104r5] as a new OGC Abstract Specification Topic to support grid-based analysis activities using various grids.

developments in the Semantic Web to link data based on geographic information in a way that provides more insight. The Location Powers: Big Linked Geodata workshop has investigated scaling effective exploitation of linked geodata by using big data approaches showing the opportunity for coordinated open developments based either on links between Big Data entities or on links between metadata for Big Data.

The OGC white paper also identifies several Open Source projects applicable to Big Geo Data under development by the Apache Software Foundation and Location Tech. Multiple members of OGC and other organizations are using them on big data applications. Parallelization is a specific area of interest.

4. OGC ACTIVITIES ON BIG GEO DATA

OGC will follow up these key actions via its main programs (innovation, standardization and outreach):

- Cloud computing for big Earth Observation data.
- Analysis Ready Data techniques including Data Cubes.
- Data Representations for Big Geo Data: Features, CRS, DGGS, Coverages.
- Linked Data applied to Big Geospatial Data.
- Using Big Data Open Source to geospatial applications.

Big Geo Data is identified as a Ripe Trend to be supported into OGC Technology Strategy. The OGC Big

Data Domain Working Group is an open forum for discussing these topics and support them across the multiple OGC Standards development related to Big Data including: WPS, WCPS, WCS, DGGS, SensorThings, Moving Features, more. A charter for a Data Cubes Domain Working Group is currently submitted to review.

The OGC Innovation Program (IP) conducts studies, pilots and other projects to advance geospatial technology innovation. These initiatives produce Engineering reports provided to the OGC Standards Program for consideration of new and refined standards. This is the case for Testbed 13 for instance and very likely for the Testbed 14. The OGC IP team also contributes on this topic to European Commission Horizon 2020 projects like DATABIO.

The OGC Communication and Outreach Program has led the Location Powers events which helped identify the topics in this white paper.

5. CONCLUSION

Interested readers are encouraged to read the full “Open Geospatial Consortium (OGC) Big Geospatial Data White Paper” [1] to benefit more deeply from all the references. They are also invited to participate to the open Geospatial Consortium works to contribute to the global efforts in support of standardization and interoperability for the benefit of all.

6. REFERENCES

- [1] [Big Geospatial Data – an OGC White Paper](http://docs.opengeospatial.org/wp/16-131r2/16-131r2.html), George Percivall, -<http://docs.opengeospatial.org/wp/16-131r2/16-131r2.html>
- [2] <http://www.locationpowers.net/pastevents/>
- [3] <http://www.locationpowers.net/pastevents/1609orlando/index.php>
- [4] <http://www.locationpowers.net/pastevents/1703delft/index.php>
- [5] <http://www.pbs.org/show/human-face-big-data/>
- [6] ISO/IEC CD2 20546 Information Technology— Big Data— Overview and Vocabulary
- [7] <https://www.iso.org/committee/45020.html>
- [8] http://dmg.org/downloads/KDD_StandardsTalkSlides/Paco.pdf
- [9] <https://github.com/SciSpark/SciSpark>
- [10] <http://www.datacube.org.au>
- [11] <http://www.earthserver.eu>
- [12] <https://www.ceosdatacube.org>
- [13] <https://www.w3.org/TR/qb4st/>

SPACE BIG DATA, SMALL EARTH LAWS OVERCOMING THE REGULATORY BARRIERS TO THE USE OF SPACE BIG DATA

Dimitra Stefoudi

Leiden University-International Institute of Air and Space Law

ABSTRACT

Space big data presents significant business, as well as scientific opportunities. Despite the substantial benefits offered, the regulatory framework around its use remains unclear. The exponential growth in the amount of data produced and disseminated consequently creates multiple legal challenges. The purpose of this paper is to briefly discuss the legal concerns stemming from the use of space big data and to propose regulatory solutions with regard to the issues of data privacy and protection, intellectual property and cybersecurity. Its aim is to analyse the legal concerns that the use of space big data is raising and suggest appropriate regulatory solutions. Towards this end, it will use the Copernicus data policy as reference and explore ways in which the said challenges can be addressed, in particular on the level of the European Union and within the framework of the general European space policy. The analysis will focus on the legal questions posed, the current regulatory regime and the way forward mainly within the scope of the European Union competences.

Index Terms— Law, data policy, data protection, cybersecurity, intellectual property, EU Law, European space policy

1. INTRODUCTION

The increasing number of launched satellites along with the need for fast and accurate information worldwide, is leading to exponential growth of the space data generated and distributed [1]. Big data from space includes not only remote sensing and Earth observation images, but also involves a great amount of data transmitted through space technology. Nevertheless, the existing regulatory framework is either partially covering or not addressing this field at all. At the moment, there is no legislation related or dedicated to space big data as a distinct field of activities. The relevant laws constitute a patchwork of international and domestic space legislation, along with regional and national data-related laws.

On EU level, policy initiatives are required in order to promote the growth of the space big data sector and promote the advancement of space applications. At the same time, there is a need for regulatory framework that would facilitate the generation, process and dissemination of space

data, according to the current technological standards and market needs.

The following sections will discuss the regulatory challenges related to space big data, focusing in particular on the European space sector, with the aim to proposing legal and policy solutions for the effective regulation and development of this field. Towards this end, issues of data privacy and protection, intellectual property, and cybersecurity connected to the use of big data from space will be addressed through the scope of EU-related legislation.

2. THE REGULATORY FRAMEWORK RELEVANT TO SPACE BIG DATA

Currently, there is no legal regime specific to the use of space big data. Space law is limited to the UN Remote Sensing Principles of 1986, which provide some general guidelines, but are of limited scope with regard to space big data. Moreover, laws on data privacy and protection, intellectual property and cybersecurity do not cover adequately the multi-faceted challenges presented. This is due to the fact that these laws were not designed to address these challenges; therefore their application on space big data uses remains ambiguous.

2.1. The Notion of Space Big Data

Space big data is a newly introduced term in legal theory and research and is used to describe large amounts of data collected by means of space technology. The large variety of available data and the predictive capabilities from its combination create the potential for well-informed decisions and accustomed services. It also offers important social benefits by establishing easy access to an unprecedented amount of information, able to address various needs of the contemporary data-driven economy and society. In this sense, big data from space is another aspect of the general big data trend and is recently growing, since more commercial actors find uses and applications for the data collected. Despite the fact that “space big data” is increasingly appearing in policy documents and scholar writings [2], there is not one official definition attributed to the term. This newly established notion is used to describe large amounts of data generated by or disseminated through space technology. The lack of definition is due to the ever developing nature of space big data, which is also the reason

why a single definition might not provide a workable solution. A term of reference though is at least necessary, in order to further define the regulatory framework relevant to space big data. This data encompasses far more than Earth observation and remote sensing imaging, given that satellite applications are nowadays used to transmit all sorts of different data. As satellite systems become increasingly connected to internet technology, space data will refer to any information going through means of space technology. Towards a more comprehensive approach of the term, identifying the constituents of space big data is essential. Among the latter are the volume in which it is generated, the velocity in which it is produced and disseminated, along with the variety and veracity of the information it contains. However, the significance of space big data is a direct outcome of the value, both scientific and commercial, that this data is able to produce. While data becomes meaningful after being combined and analysed, it only obtains certain value when it is exchanged among interested parties. The growing demand for open data, in conjunction with the increasing accuracy of generated data, poses challenges beyond the current regulatory framework. Space big data is a topical issue, interesting from the point of view of business opportunities in the space and other sectors, scientific advancement on Earth and in outer space, as well as overall improvement of daily lives with the use of space technology.

2.2. Legal challenges & EU laws

On the one hand, securing a clear and stable regulatory regime for the generation and use of space data is a crucial step towards incentivising its users and maximising the benefits offered from space data applications. On the other hand, it is essential to ensure a minimum degree of compliance with regulatory restrictions. In this context, the debate over privacy versus openness forms the basis of the regulatory conundrum. Therefore, the importance of understanding and addressing the legal concerns surrounding the use of space big data is becoming topical.

The most prominent challenge is related to data privacy and protection. In addition, intellectual property rights become redundant, since it is not feasible to track the creator of the protected property or the ways in which it is distributed. What is more, the issue of cybersecurity lately also appears to be alarming satellite operators. Satellite and ground stations alike contain large datasets that could potentially be exposed to external risks.

2.2.1. Data Privacy & Data Protection

The relevance of data privacy and protection with regard to the use of space big data is twofold. On a first level, the increasing quality of Earth observation images is posing obvious concerns, especially given the growing commercial market of high-resolution imaging. On a second level, the growing number of throughput satellites combined with

increasing reliance on satellite technology for connectivity services extends the types of space data to any information that is shared through this means.

The rapid cross-border data dissemination does not allow for a specific compliance regime to apply. Satellite signals, for example, are transmitted within fractions of seconds among multiple satellites in orbit, ground stations, databases, and all sorts of electronic devices. The variety of data available and the speed, in which they are transmitted, along with the numerous ways of processing, create a situation where the data subject, the data analyst and the final product are hard to distinguish and locate. Consequently, existing regulations are either partially or inadequately addressing the challenges.

In the European framework, the recent General Data Protection Directive [3] is imposing even stricter limits to data sharing and processing, while it also puts restraints to data access and transfer and stretches its scope of application outside the EU borders. At the same time, various different data privacy and protection regulations in EU member States and third countries compile a complex regulatory regime.

2.2.2. Intellectual Property Rights

Intellectual property is another issue connected to the use of space big data to the extent that the latter is processed and analysed after its collection and before its distribution. Intellectual property rights refer to the entitlement of the author of an original work to control the reproduction of the said result. It usually appears in the form of copyright or patent protection and is heavily regulated on international, European and national level [4]. In terms of space big data, this right is mainly relevant with respect to Earth or space observation images, as well as to databases and data processing software [5]. However, given the rapid distribution of space big data and its various intermediate uses, tracing the creator of this data and its subsequent users is becoming cumbersome.

2.2.3. Cybersecurity

In view of the aforementioned challenges of data privacy and intellectual property, the purpose of cybersecurity is to protect these legitimate rights. Cybersecurity threats are linked to the vulnerability of datasets to external and unauthorised attacks, mostly carried out through access to computer systems and interference with the stored data [6]. Such threats are considerable concern for satellites and other connected systems, given their growing integration with internet technology and the transmission of various sorts of data [7]. The relevance of cybersecurity to the space sector is related to databases and interference with satellite systems [8].

There are several regulations that provide for standards to tackle cyber-attacks [9], while the issue is currently discussed on within the scope of EU Law [10]. In terms of

the latter, priority is given to the formulation of a comprehensive set of measures and requirements to render systems less vulnerable, as well as on ways of efficient incorporation into practice.

Some of the aforementioned challenges are addressed within the scope of the Copernicus Regulation and the prescribed Data Policy. Even though the Regulation creates a balance between open data and the said concerns, its scope is limited to Earth observation. Nevertheless, some of its elements can influence the discussion on future regulation in the field of space big data.

3. COPERNICUS: LESSONS LEARNT

Recognising the added value of space data and its potential to create scientific and business opportunities, as well as to encourage entrepreneurship in Europe and worldwide, the European Union initiated the Copernicus flagship programme, the world's largest single Earth observation mission. The Copernicus policy is based on “full, free and open” data from space. Copernicus is one of the first instances where the term space big data appeared and revealed the potential of space big data applications.

In 2014 the European Commission adopted the Copernicus Regulation [11], to govern one of its two flagship programmes, which is owned by the European Union and procured by the European Space Agency. The purpose of the programme is to offer access to a large pool of Earth observation data, mostly of low and medium resolution, collected by its Sentinel satellites, as well as by the Copernicus Contributing Missions. The Copernicus Regulation dedicates a separate title to “Data Policy and Security”. The main feature of Copernicus, described in this section, is the open access to its data with no discrimination on the basis of the origin and purposes of the users [12]. However, this seemingly unrestricted access can be limited on the grounds of security concerns [13], which is mostly relevant to higher resolution or other sensitive data. This way, the Copernicus Regulation managed to achieve a balance between data openness and data protection, setting some safeguards towards the secure use of its services by the subscribers. Its provisions could serve as a basis for further regulation of the use of space big data in Europe.

4. THE WAY FORWARD: EU LAW AND POLICY

The most important step forward is to recognise the significance of space big data, its uses and applications, as an individual issue of space policy on European level. This would pave the way for the desired uniformity, which will further facilitate productive efforts towards regulating the field of space big data.

Even though the European Union has recognised the fundamental character of data uses for economic and social growth [14], it is lacking sector-specific policies that focus on space data. Policy initiatives recognising the benefits of space big data will enable the identification of the particular

needs and challenges, and promote its use, by creating a bridge for the use of space big data applications outside the space sector. To this end, and in order to ensure a comprehensive approach to the subject, dialogue and cooperation among the stakeholders is essential in voicing the challenges faced and the issues to be addressed.

In addition to policy matters, regulatory requirements with regard to the use of space big data are becoming increasingly imminent. The existing legal regime is either insufficient or not directly addressing the aforementioned challenges posed by big data from space. However, before proceeding with regulatory endeavours, the fast-paced technological development in this field should be taken into account. On the one hand, a certain degree of legal certainty will facilitate the use of space data. On the other hand, an irrelevant or exceedingly restrictive regulation may pose impediments in this regard. A flexible and adaptive approach should be able to connect these two ends in a balanced way.

Primarily, regulatory focus should be put on providing comprehensive definitions of crucial notions, which are needed for further defining the laws applicable on space big data. This would consequently clarify the part of existing legislation relevant to the use of space data and help assess whether further regulation is needed.

In order to overcome the risk of strict regulation that will be outgrown by the rapid technological developments, regulatory attempts should focus on the formulation of standards and recommended practices. The latter would reflect the actual practice in the field of space big data applications and be able to cope up with the fast technological advancement. This way, the need for a flexible legal regime will be summarised in a set of minimum standards that will serve as benchmark for compliance with other relevant regulatory requirements.

Apart from the need for an adaptable regime, uniformity is another important factor influencing future regulatory initiatives. A streamlined framework on the level of the European Union and its member States will provide greater legal certainty, a competitive advantage in favour of the European industry. Furthermore, legislation on European level, even in the form of softer law and not necessarily a directive or regulation provides a feasible solution, as it is less complex to adopt and update.

Following the Copernicus example, it is important to distinguish between data categories that for various reasons require regulation and others that could be made freely available. As a result, regulatory constraints will be imposed only to the extent that they are required, while regulatory freedom will be ensured, in order to promote and increase the use of space big data [15].

5. CONCLUSION

The importance of the use of space big data was already acknowledged in the promulgation of the Copernicus EU

flagship programme [16]. The growing scientific and commercial interest in the use of big data from space is confirming its beneficial character, but also poses challenges to the existing regulatory regime. As technology progresses, regulation should follow in a way that will complement and facilitate further development. This can be achieved through concerted efforts on European level in response to the growing concerns.

The creation of a comprehensive policy and regulatory environment for the use of space big data will promote the competitiveness of the European Union in this field and attract innovation and growth opportunities.

6. REFERENCES

- [1] K. Russell, Satellite Launches to Increase Threefold Over the Next Decade By Kendall Russell, Via Satellite, October 12 2017, <http://www.satellitetoday.com/newspage/2017/10/12/satellite-launches-increase-threefold-next-decade/>
- [2] A comprehensive definition that combines elements mentioned in various documents is provided on the Copernicus website, with reference to the ESA BiDS conference. Space big data is described as “massive spatio-temporal earth and space observation data collected from space-borne and ground-based sensors”, <http://www.copernicus.eu/events/big-data-space>
- [3] Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 4.5.2016, L119, 2016
- [4] Article II.viii (content of intellectual property rights), WIPO Copyright Treaty, 36 ILM 65, 1997; EU Directive 96/9/EC on the legal protection of databases, OJ 27.3.1996, L77, 1996; EC Communication, A Single Market for Intellectual Property Rights, 24.5.2011, COM(2011) 287 final
- [5] F. von Der Dunk, Earth Observation Data Policy in Europe: An Inventory of Legal Aspects and Legal Issues in R. Harris, Earth Observation Data Policy and Europe, 2002, 26
- [6] M. Maybaum, Technical Methods, Techniques, Tools and Effects of Cyber Operations in K. Ziolkowski (ed.), Peacetime Regime for State Activities in Cyberspace. International Law, International Relations and Diplomacy, NATO CCD COE Publication, Tallinn 2013, 103.
- [7] P.J. Blount, Satellites are just things on the Internet of Things, 42.3 Air and Space Law 274, 2017, 279 and 288
- [8] S. Kaiser, M. Mejia-Kaiser, Cybersecurity in air and space law, 64 German Journal of Space Law 396, 2015, 398 and 404; Committee on the Peaceful Uses of Outer Space’s (COPUOS) guidelines for the long-term sustainability of outer space activities, A/AC.105/C.1/L.354, 2016, Part B, Guidelines 9, 18 and 19
- [9] Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace, 7.2.2013, JOIN(2013) 1 final; The European Agenda on Security, 28.4.2015, COM(2015) 185 final; Directive concerning measures for a high common level of security of network and information systems across the Union, OJ 19.7.2016, L 194, 2016
- [10] Proposal for a Regulation on ENISA, the “EU Cybersecurity Agency”, and repealing Regulation 526/2013, and on Information and Communication Technology cybersecurity certification (Cybersecurity Act), 4.10.2017, COM(2017) 477 Final2; Joint Communication to the European Parliament and the Council, Resilience, Deterrence and Defence: Building strong cybersecurity for the EU, 13.9.2017, JOIN(2017) 450 final
- [11] EU Regulation 377/2014 establishing the Copernicus Programme and repealing EU Regulation 911/2010, OJ 24.4.2014, L 122/44
- [12] Article 23.2 of the Copernicus Regulation
- [13] Articles 24 and 25 of the Copernicus Regulation
- [14] Communication on the Mid-Term Review on the implementation of the Digital Single Market Strategy, A Connected Digital Single Market for All, 10.5.2017, COM(2017) 228 final; Commission Staff Working Document on the free flow of data and emerging issues of the European Data Economy, 10.1.2017, SWD(2017) 2 final; European Commission Communication, Space Strategy for Europe, 26.10.2016, COM(2016) 705 final, 2-3
- [15] M. C. Hansen, T.R. Loveland, A review of large area monitoring of land cover change using Landsat data, 122 Remote Sensing of Environment 66, 2012, 71
- [16] Study to examine the socio-economic impact of Copernicus in the EU, Report on the socio-economic impact of the Copernicus programme, PwC, October 2016, http://www.copernicus.eu/sites/default/files/library/Copernicus_SocioEconomic_Impact_October_2016.pdf

STARE - TOWARD UNPRECEDENTED GEO-DATA INTEROPERABILITY

Kwo-Sen Kuo^{1,2}, Michael L. Rilee^{1,3}

¹NASA Goddard Space Flight Center, Greenbelt, Maryland, USA;

²Bayesics LLC, Bowie, Maryland, USA;

³Rilee Systems Technologies LLC, Derwood, Maryland, USA

ABSTRACT

As a universal geoscience data representation, the Spatio-Temporal Adaptive-Resolution Encoding, STARE, is bringing about unprecedented interoperability to all Earth Science data. In its spatial component, STARE contracts the usual two-dimensional, i.e. latitude and longitude, geolocation into a one-dimensional, hierarchical index. The STARE geolocation index follows the quadfurcation scheme of the well-established hierarchical triangular mesh (HTM) used in astronomy but with an innovative bit-field arrangement that includes approximate data resolution information to enable efficient geospatial set operations. STARE's temporal component is also hierarchical with bit fields referring to conventional date-time intervals or units. STARE is designed for geo-spatiotemporal data placement alignment in databases (e.g. SciDB) but also supports more traditional contexts via a STARE application programming interface (API).

Index Terms— interoperability, interdisciplinary analysis, universal data representation, geo-spatiotemporal indexing, data placement alignment, array database

1. INTRODUCTION

Facing the deluge of ever increasing volumes of data, there have been many efforts attempting to address the growing challenge of Earth Science data practice with Big Data technologies. Most of these efforts emphasize the volume aspect of the challenge. We, however, have recognized variety as the key [1] to attaining optimal scientific value.

Parallelization is the obvious solution to the volume challenge. But, without variety homogenization, the scalability of parallelization is at best piecemeal, i.e. one variety at a time. Since Earth Science is a system science, its investigations often require integrative, interdisciplinary analyses that use multiple, diverse datasets. Thus, a complete solution needs to effectively address the variety challenge.

1.1. Sources of Variety

Although there are different kinds and aspects of variety, we have focused on what we believe to be the most crucial to Earth Science data in terms of data-analysis efficiency gain,

i.e. the diversity in data model. While the fundamental data structure remains the same, i.e. array, there are three data models used by NASA Earth Observing System Data Information System (EOSDIS) for its data collection: Grid, Swath, and Point, embodied in the HDFEOS [2] software library based on the hierarchical data format (HDF). Each data model is designed for a specific class (or level) of data products. The underlying variety, however, is significantly more diverse than that suggested by the three data models.

The detailed characteristics of how the data are obtained and processed produce more varieties and thus drastically complicate the matter, including differences in satellite orbits, viewing strategies, modes of operation, spatial resolutions of the instruments, etc. The resultant diversity seriously hinders data interoperability. Consequently, finding a universal data representation that can homogenize them all presents a formidable challenge.

2. DESCRIPTION OF STARE

In our effort to identify the likeliest approach to achieve optimal value when dealing with Big Earth Data using a distributed parallel array database (i.e. SciDB), we come to realize that a better indexing scheme for geolocation is needed and critical. This is because SciDB uses array index to control data partitioning and hence placements of partitioned data chunks.

The primary requirements for such an indexing scheme are: 1) It needs to support spatiotemporal data placement alignment for array databases; and 2) It needs to include resolution information of the underlying data.

The rationale for the first requirement is straightforward: Most integrative analyses in Earth Science require spatiotemporal coincidence. Data placements aligned spatiotemporally thus ensure the minimization of node-to-node communication on a distributed parallel database and, as a result, performance optimization. However, due to the diversity in data models, as well as in data products and data granules, no unique and uniform correspondence exists between geolocation and data array index. That is, the same array index in different data granules often refer to different geolocations. This inconsistency represents the greatest difficulty to satisfy the first requirement, especially if we wish to meet Big Data challenges with systematicness and automation.

The second requirement is to ensure that set operations for integrative analyses over multiple, diverse datasets can be carried out robustly and consistently without sacrificing performance. While there may be many geo-indexing schemes satisfying the first requirement, few can simultaneously satisfy this second requirement.

The SpatioTemporal Adaptive-Resolution Encoding, STARE, is the innovative outcome of much research and deliberation that satisfies both requirements. It consists of two parts, a spatial and a temporal component as described in the subsections below.

2.1. Spatial component

Hierarchical triangular mesh (HTM) [3][4] is a way to address the surface of a sphere (or, more accurately, the solid angle) using a hierarchy of spherical triangles. The mesh is generated with the procedure below:

1. Start with an inscribing octahedron of a sphere.
2. Bisect each edge of the triangular facets.
3. Bring the bisecting points to inscribe the sphere to form 4 smaller spherical triangles.
4. Repeat from step 2, until a desired resolution (uncertainty) is reached.

After the initial octahedron, each iteration from step 2 is termed a *quadfurcation*, i.e. division/branching into 4 parts.

The spatial component of STARE is a customized variant of the HTM with two distinctions. 1) While right-justified encoding is used for the original HTM indexing, we choose a left-justified encoding to facilitate spatial data placement alignment. 2) In addition, geolocation uncertainty (commensurate with data resolution) is added to the encoding using a few least-significant bits to facilitate set operations among diverse datasets [1].

Essentially, STARE's spatial index is a one-dimensional equivalent way (to the use of latitude-longitude) of specifying geolocation to a given uncertainty. For example, with 23 quadfurcations (i.e. at the 23rd depth level), a latitude-longitude coordinate is concisely and uniquely mapped to an integer with ~1-m uncertainty. Table 1 shows the uncertainty at each level of quadfurcation in detail.

2.2. Temporal component

STARE's temporal component is also hierarchical but uses calendrical date/time units to avoid unnecessary translations between temporal frameworks. Table 2 provides just one example encoding with a maximum time resolution of milliseconds, common for observations obtained from spacecraft. As an example, for an observation made on 2015 June 12, 8:10 AM with millisecond resolution, STARE yields

[+] 000-002015-06-3-3 08:0600.000 (07).

Here, the '(07)' corresponds to the highest level (finest) resolution available, milliseconds, and the '['+' signifies "positive" years. With the Unix-based convention of num-

D	R	L	D	R	L
23	~1 m	~1.2 m	11	~4 km	~5 km
22	~2 m	~2.4 m	10	~8 km	~10 km
21	~4 m	~5 m	9	~16 km	~20 km
20	~8 m	~10 m	8	~31 km	~39 km
19	~15 m	~19 m	7	~63 km	~78 km
18	~31 m	~38 m	6	~125 km	~157 km
17	~61 m	~77 m	5	~251 km	~314 km
16	~122 m	~153 m	4	~501 km	~628 km
15	~245 m	~307 m	3	~1003 km	~1,256 km
14	~490 m	~615 m	2	~2005 km	~2,500 km
13	~1 km	~1.2 km	1	~4011 km	~5,000 km
12	~2 km	~2.5 km	0	~8021 km	~10,000 km

Table 1 Approximate uncertainties in radius (**R**) of the area and the edge length of the triangle (**L**) at each level of HTM quadfurcation starting with the octahedron at depth (**D**) 0.

bering months starting at 0 and days at 1, the native STARE format can be read (partly) as the 3rd day of week 3 of month 6 (the 7th regularized 28-day month). Conversion to an array index is simple, yielding

x407b6d04b0007,

and alternative encodings may be devised to meet application requirements.

Range	Starting Bit	Ending Bit	No. Bits	Meaning
0	0	2	3	Resolution/Unit
1	3	12	10	millisecond
2	13	24	12	Second
3	25	29	5	Hour
4	30	32	3	Day of week
5	33	34	2	Week
6	35	38	4	Month
7	39	48	10	Year
8	49	58	10	Kilo-annum
9	59	62	4	Mega-annum
10	63	63	1	Before/After

Table 2 An example STARE temporal encoding.

3. ADVANTAGES OF STARE

Since STARE indexing is hierarchical and carries with it (approximate) resolution information, it embodies many advantages that are hard to find all together in an alternative. We list some of the most important ones below.

First, STARE affords sophisticated and yet highly efficient set operations, including *conditional subsetting*, which samples a second dataset based on properties (usually filtered) of a first dataset. For example, one may wish to correlate cloud-top infrared brightness temperature with the presence and intensity of precipitation. Such an analysis may thus lead to obtaining the brightness temperature at 8.6-

micron from MODIS¹ where and when a TRMM² orbit granule indicates presence of precipitation. With STARE, this sort of set operations is turned into fast operations on STARE-index integer intervals, as opposed to much slower operations on floating-point latitude-longitude pairs.

Figure 1 shows an example of the spatiotemporal overlap of a TRMM orbit swath (black strip with shades of blue indicating precipitation intensity) and a Terra/MODIS 5-min granule (brightness temperature at band 29 in shades of cyan and yellow). For presentation clarity, coarse-resolution coverings of HTM spherical triangles are overlaid on both the TRMM swath (blue triangles) and the MODIS granule (red triangles). Higher-resolution triangles will not only make the image too busy but also the list in Table 3 too long, which shows, in hexadecimal, the STARE indices of the triangles for the HTM covering of the MODIS granule.

Moreover, because each edge of a spherical triangle in HTM is a segment of a great circle, it is more straightforward to ascertain which hemisphere (delineated by the great circle) a given geolocation belongs to. This property can thus be utilized to quickly determine the set of STARE indices (even with varying quadfurcation levels) corresponding to a user specified region of interest (ROI).

The hierarchical nature of STARE also supports progressive visualization. That is, we may use coarser resolution (lower-level quadfurcation) and thus smaller data volume to rapidly render initial visualization and use progressively higher resolutions to refine the visualization until a desired quality is reached. Bandwidths in a data traffic chain, especially when low-bandwidth connections (e.g. Internet) are involved, can therefore be better utilized to provide a more pleasant user experience.

A *disadvantage* of STARE, however, is the lack of a straightforward way in specifying an overlap (aka halo) for

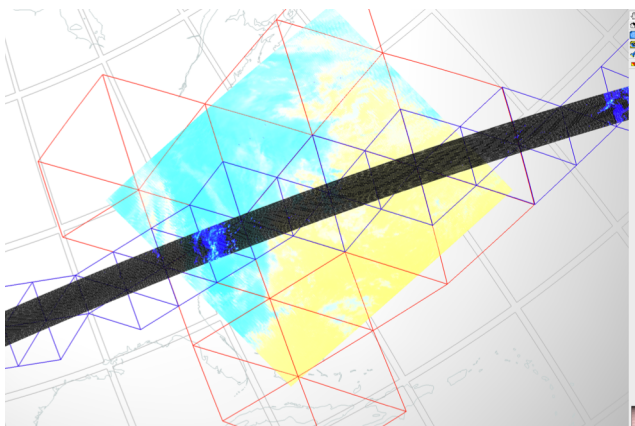


Figure 1 Overlap of a TRMM orbit swath with a Terra/MODIS 5-min granule with coarser-resolution covering HTM spherical triangles overlay. See text for details.

¹ MODIS - MODerate-resolution Imaging Spectroradiometer, an instrument on the Terra and Aqua platforms of [NASA's Earth Observing System](#) (EOS).

² TRMM - NASA's [Tropical Rainfall Measuring Mission](#).

neighboring partitions of data, which is important to performance for operations that are not pleasingly parallel. Determining neighboring STARE cells is straightforward, but requires a (small) tree traversal, as opposed to a simple index increment, that must then be convolved with the parallel computing platform's data distribution scheme.

4. IMPLEMENTATIONS OF STARE

We use STARE to support our research into the automated analysis of phenomena such as blizzards as moving objects. We are extending the capabilities of the array database SciDB and developing tools (database ingest, preprocessing, visualization) using STARE through a software library and application programming interface (API).

SciDB is an array database allowing scientists to analyze voluminous datasets as arrays with little concern for the details as to how it achieves massively parallel processing. STARE adds a spatiotemporal interpretation to SciDB. Data partitions based on STARE indexing are now naturally collocated in time and space when distributed across parallel computing and storage resources. We have found STARE's robust combination of geometry and array indexing invaluable for adding Climate Data Operators [5] based regridding as a User Defined Operators (UDOs) in SciDB. STARE aids the sorting and finding of relevant data points that are key to interpolating between grids, even in serial applications.

Using STARE to organize data before database ingest, we can scalably send data chunks directly to the correct node, avoiding costly data transfers and repartitioning. With

```
<notional-modis-stare-metadata>
<TemporalIndex>
  x404e0052c0007 x404e005517fff
</TemporalIndex>
<SpatialIndex>
  x4a00000000000004,
  x4a20000000000004 x4a3fffffffffffffff,
  x4a50000000000004,
  x4a80000000000004 x4adfffffffffffffff,
  x4af000000000004,
  x4b1000000000004,
  x4b60000000000004,
  x4be000000000004,
  x4c00000000000004 x4c1fffffffffffffff,
  x4c30000000000004 x4c7fffffffffffffff,
  x4ca0000000000004,
  x4cc0000000000004,
  x4ce0000000000004 x4cfffffffffffffff,
  x4f80000000000004,
  x4fa0000000000004 x4fbfffffffffffffff
</SpatialIndex>
</notional-modis-stare-metadata>
```

Table 3 Example text-representation of STARE metadata for Terra/MODIS granule MOD021KM.A2009337.0240 corresponding to the red triangles in **Figure 1**. Logical operations such as intersection and conditional subsetting become operations on integer intervals like these.

the enormous scale of Earth Science datasets, the savings from eliminating costly data repartitioning cannot be over-emphasized.

5. STAND-ALONE STARE

While the strengths and benefits of STARE are best exploited and manifested in a distributed parallel array database system like SciDB, such systems are not necessary in order to enjoy a subset of its benefits. For example, it is mentioned above that STARE helps us in the implementation of regridding functions. In the following two subsections, we outline respectively two envisaged adaptations of STARE to existing data practice for immediate interoperability improvement in more traditional contexts.

5.1. As an enabling companion to existing data files

One way that existing data practice may take advantage of STARE right away is by augmenting each existing data file with an enabling helper companion in the same format as the data file. This helper companion effectively contains a forward lookup table and a reverse one: 1) a forward table for looking up STARE indices using the indices of the data array contained in the data file and 2) a reverse table for looking up data array indices using STARE indices.

With the lookup tables as helper companions, more streamlined and sophisticated subsetting, e.g., conditional subsetting, can be performed among diverse datasets with existing data practice, albeit in a less direct manner than with SciDB. A typical usage example of finding environmental conditions for where and when a Level 2 (L2) TRMM Swath dataset indicates precipitation is likely to follow these steps:

1. Use a L2 TRMM dataset to identify precipitation regions;
2. Convert the array indices of the precipitating instantaneous fields of view (IFOVs) to STARE indices using the forward lookup table;
3. Find intersects with, say, MERRA-2³, data products using the STARE indices associated with MERRA-2 data files;
4. Convert the STARE indices to MERRA-2 array indices using the reverse lookup table; and
5. Obtain environmental conditions, e.g. pressure, temperature, humidity, etc., from MERRA-2 for the precipitating regions indicated by TRMM.

It is conceivable that these helper lookup tables may be adopted as standard components of standard data files (rather than in separate files) after they have proven their worth.

³ MERRA - Modern Era Retrospective-analysis for Research and Application, NASA's premier reanalysis data product.

5.2. As geo-spatiotemporal metadata

Remembering that STARE is hierarchical, it stands to reason that there is no need to keep STARE indices of all IFOVs or grid cells if the objective is to specify the spatio-temporal extent covered by the data in a file. Instead, larger spherical triangles and temporal intervals can be used, which will result in a smaller number of STARE indices and reduced data volume, making it suitable for inclusion as metadata, e.g. see Table 3 for a dramatic example using a fairly coarse STARE resolution to support approximate set functions. Such geo-spatiotemporal metadata may be used to enable more (but basic) subsetting options at the data discovery stage, for example, finding spatial (and/or temporal) overlaps between data granules of diverse datasets.

6. SUMMARY

STARE is a flexible scheme for encoding geo-spatiotemporal information, providing a universal geolocation index for Earth Science Data. When used as an indexing scheme, STARE helps automate the scalable use of distributed, massively parallel compute and storage resources to manage and analyze Big Earth data. We have developed a basic STARE capability and used it within the framework of SciDB array database as well as more conventional data processing tools (e.g. via the command line or scientific visualization). STARE proved extremely valuable as we added regridding to the SciDB array database and expect it to be critical for the scalable fusion and integration of diverse, massive Earth Science Data.

Acknowledgement - This research is primarily supported by the NASA Earth Science Technology Office (ESTO) through its Advanced Information Systems Technology (AIST) Program. Supplementary support is provided by the NASA Earth Science Data Systems program through its Advancing Collaborative Connections for Earth System Science (ACCESS) program

REFERENCES

- [1] Rilee, M. L., K-S Kuo, T. Clune, A. Oloso, P. G. Brown, and H. Yu, "Addressing the big-earth-data variety challenge with the hierarchical triangular mesh," *2016 IEEE International Conference on Big Data (Big Data, IEEE)*, 1006–1011, 2016.
- [2] HDF-EOS5 Data Model, File Format and Library: <https://earthdata.nasa.gov/standards/hdf-eos5>
- [3] P.Z. Kunszt, A.S. Szalay, and A.R. Thakar, "The Hierarchical Triangular Mesh. In Mining the Sky" Proceedings of the MPA/ESO/MPE Workshop, Garching, Berlin/Heidelberg, Ch. 83, p631, 2001.
- [4] A.S. Szalay, J. Gray, G. Fekete, P.Z. Kunszt, P. Kukol, and A. Thakar, "Indexing the Sphere with the Hierarchical Triangular Mesh," *Micr. Res. Tech. Rpt., MSR-TR-2005-123*, 2005.
- [5] Climate Data Operators: <https://code.mpimet.mpg.de/projects/cdo>

CREATING VIRTUAL SEMANTIC GRAPHS ON TOP OF BIG DATA FROM SPACE

Konstantina Bereta and Manolis Koubarakis

National and Kapodistrian University of Athens

ABSTRACT

We present the system Ontop-spatial for the integration of geospatial data from different sources and different formats using ontologies and mappings. Ontop-spatial answers GeoSPARQL queries over geospatial relational databases storing vector or raster data by performing on-the-fly GeoSPARQL-to-SQL translation using ontologies and mappings. Our experimental evaluation shows that Ontop-spatial outperforms all state-of-the-art geospatial RDF stores.

Index Terms— GeoSPARQL, RDF, Ontology-based Data Access, Geospatial Data Integration

1. INTRODUCTION

Previous projects TELEIOS, LEO and Melodies funded by FP7 ICT, and OBEOS funded by ESA have demonstrated the use of linked data in Earth Observation (EO). The current H2020 project Copernicus App Lab (<http://www.app-lab.eu/>) goes one step further by making data from three Copernicus services (Land, Marine and Atmosphere) available on the Web as linked data to aid their utilization by mobile developers. In previous projects, it has been assumed that EO data are transformed from their original formats (shapefiles, spatially-enabled relational databases, GeoTIFF, NetCDF etc.) into RDF, stored in geospatial RDF stores and queried using geospatial extensions of SPARQL to develop interesting applications. In this paper, we present the system Ontop-spatial which enables the creation of *virtual RDF graphs over EO data stored in their original formats* using ontologies and mappings. Ontop-spatial allows EO data centers to make their data available as linked data that can be queried using the OGC standard GeoSPARQL [1], without first having to translate this data into RDF. Ontop-spatial scales to big geospatial data and it is more efficient than related geospatial RDF stores. Ontop-spatial is available as open source at <https://ontop-spatial.di.uoa.gr>.

Ontop-spatial adopts the Ontology-Based Data Access (OBDA) paradigm pioneered by the Semantic Web community, and it is the *first geospatial OBDA system*. Ontop-

This work has been funded by the EU project Copernicus App Lab (730124).

spatial is able to connect to geospatial databases and create geospatial RDF graphs on top of them, using ontologies (that are extensions of the GeoSPARQL ontology) and R2RML mappings. Figure 1 shows graphically the classes of the GeoSPARQL ontology.

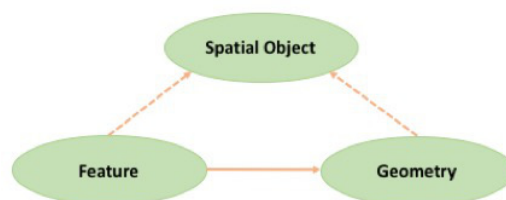


Fig. 1. Overview of evaluation results

R2RML (<https://www.w3.org/TR/r2rml/>) is the standard language for encoding how relational data is mapped into RDF terms. This virtual approach avoids the need of materialization and facilitates data integration, as it enables users to pose the same GeoSPARQL queries they would pose over the materialized RDF data. GeoSPARQL queries are translated by Ontop-spatial on-the-fly into the respective SQL queries with spatial operators, and are evaluated in the geospatial DBMS. Currently, PostGIS, Spatialite and Oracle Spatial are supported as back-ends. The first version of Ontop-spatial dealing with vector data only has been presented in [2]. This version has been used in three environmental applications in the context of project MELODIES, and in a marine-security application in the context of German national project EMSec.

The contribution of our approach with respect to the Big Data from space dimensions are the following. **Volume:** Ontop-spatial outperforms the state-of-the-art in geospatial RDF stores and is able to process tens of Gigabytes of data containing complicated geometries. **Velocity:** When data gets frequently updated, using traditional triple stores is inefficient, as batches of data need to be converted and materialized as RDF triples each time they arrive. Our approach eliminates as much as possible the need for materializing data and it is suitable for data sources that get frequently updated (e.g., streams). **Variety:** With raster and OPenDAP support in place, Ontop-spatial becomes the first GeoSPARQL query engine that is able to process such a wide variety of geospatial

formats, enabling geospatial data integration using ontologies and mappings. **Value:** Exposing geospatial data as virtual RDF triples that can be accessed in the Web through standard (Geo)SPARQL endpoints enables the interlinking of EO data with other data (e.g., open data) increasing their value, as data from multiple geospatial sources can be combined and rich queries can be expressed over them.

2. ONTOP-SPATIAL

Since the publication of [2] which discussed how Ontop-spatial can be used to query vector data, we have extended Ontop-spatial with the ability to query raster data as well. Querying raster data sources using declarative query languages can also be done using array DBMSs such as Rasdaman, MonetDB and SciDB. As GeoSPARQL does not include support for raster data, in our approach we do not deviate from the standard but instead: i) we overload existing vector GeoSPARQL operators such as `geof:sfIntersects` to be used with raster data as well, and ii) in the mappings, we use the raster functions supported by the underlying DBMS (e.g., PostGIS with the raster support).

More recently, work on the SciSPARQL query language showed how to query grid coverages using a hybrid data store composed of Rasdaman and a main-memory RDF store [3]. We deviate from this approach by not extending (Geo)SPARQL with array functionalities but allowing for the encapsulation of raster data functions in the mappings, so that not every raster cell needs to be represented in RDF.

The problem of representing and querying raster data as linked data has also been discussed in the recent working note "Coverages in Linked Data" by the OGC/W3C Spatial Data on the Web working group (https://www.w3.org/2015/spatial/wiki/Coverages_in_Linked_Data).

None of the geospatial extensions of the framework of RDF and SPARQL, such as stRDF and stSPARQL and GeoSPARQL have considered support for raster data. The main challenge that lies behind this is twofold. First, a raster file is associated with a geometry only as a whole. It is not straightforward to associate separate raster cells to a geometry; they have to be vectorized first (i.e., translated into polygons). Second, every raster cell is associated with one or more values. In order to convert all information contained in a raster file into RDF, then multiple triples should describe a raster cell, producing a large amount of triples for a whole raster file. However, not all of this information is needed. In most of the use cases, only the information that derives from a raster file and satisfies certain criteria (e.g., value constraints) is all that is needed to be converted into RDF. This means that the raster file needs to be processed and then the results of this processing are useful as RDF, while any other information is redundant. These challenges have discouraged the scientific community from converting and materializing raster data to

RDF. The following example describes how raster data can be mapped into virtual RDF data. For the convenience of the reader, we present the mappings using the OBDA native language of Ontop instead of R2RML, as it is more compact and readable, but R2RML is also supported in the system.

```
mappingId chicago2
target geo:{geom} rdf:type f:rastCell;
      geo:asWKT {geom} .
source  select ST_DumpAsPolygons(rast)
as geom from chicago;
```

In the example described above, a GeoTIFF image has been imported into a PostGIS database as relation `chicago`. The mapping shows how raster data stored in column `rast` are mapped to geometries in WKT format, after they are vectorized, using the PostGIS `ST_DumpAsPolygons` function. This is a procedure that allows domain experts to use all geometries that they may have in a database uniformly, and execute spatial operations involving vector and raster geometries. Domain experts usually perform this vectorization step as part of pre-processing. In the mapping described above, we show how this can be done on-the-fly, using Ontop-spatial.

In the project Copernicus App Lab, Ontop-spatial has also been extended to support data sources made available via OPeNDAP services offered by our partner Dutch company RAMANI. OPeNDAP is a framework for accessing scientific data (<https://www.opendap.org/>) which is widely used by Earth scientists, as it is popular in large organizations such as NASA and NOAA. Earth science data can be consumed by using a specific OPeNDAP client. To make data provided by OPeNDAP services available as linked data, the data should be downloaded, materialized and then converted into RDF using custom programs, as existing applications that convert geospatial data into RDF do not offer support for OPeNDAP. The approach that we describe in this paper enables the creation of virtual geospatial RDF graphs on top of data that is accessible through OPeNDAP on-the-fly, without materializing the original data or the RDF data.

Ontop-spatial has been extended with an adapter that enables it to retrieve data from an OPeNDAP server, create a table view on-the-fly, populate it with this data and create virtual semantic geospatial graphs over it. To achieve this, Ontop-spatial utilizes the system MadIS (<https://github.com/madgik/madis>) as a back-end. MadIS is an extensible relational database system built on top of the APSW SQLite wrapper. MadIS is a framework that provides a Python interface so that users can easily implement user-defined functions (UDFs) as row, aggregate functions, or virtual tables. We used MadIS in order to create a new UDF, named `OpEndap`, that is able to create and populate a virtual table on-the-fly with data retrieved from an OPeNDAP server. In this way, Ontop-spatial enables users to pose GeoSPARQL queries on top of OPeNDAP data sources without materializing any triples or tables.

An example is provided below.

```

mappingId opendap_mapping
target lai:{id} rdf:type lai:Observation ;
      lai:{id} lai:hasLai {LAI}^^xsd:float;
      lai:detectionTime {time}^^xsd:dateTime;
      geosparql:asWKT {wkt}^^geo:wktLiteral .
source select id, LAI, time, wkt
      from (ordered opendap
      url:https://analytics.ramani.ujuizi.com/
      %28https://ramani.ujuizi.com/
      thredds/dodsC/Copernicus-Land-timeseries-
      global-LAI%29/readdods/.LAI/)
      where LAI > 0

```

In this mapping, the `source` is a Leaf Area Index (LAI) dataset with resolution of 100 meters is provided through an OPeNDAP server. The dataset contains observations about the LAI values of areas, as well as the time and location for each observation. The MadIS operator `Opendap` retrieves this data and populates a virtual database table with the schema `(id, LAI, time, wkt)`. The column `id` was not originally in the dataset but it is constructed from the location and time when the observation is taken. The LAI column stores the LAI values of an observation as `float` values. The attribute `time` represents the timestamp of an observation in `datetime` format. In the original dataset temporal values are represented as numeric values. The meaning of these values is described in the metadata. For example, it can be days or months since a fixed timestamp. Unfortunately, this is not a standard representation that we would have available if we had imported the dataset into a geospatial database. Because of the fact that the `Opendap` operator is implemented as an SQL user-defined operator, it can be embedded into any SQL query. So we refined the data that we want to be translated into virtual RDF terms by adding an SQL filter to the query to eliminate the negative or zero LAI values by filtering them out at an intermediate level, so that i) we do not change the values of the original dataset and ii) we provide only the correct values to the users so that they do not need to handle the noise themselves (e.g., by using GeoSPARQL filters or custom code).

The `target` part of the mapping encodes how the relational data can now be mapped into RDF terms. Every row in the virtual table describes an instance of the class `lai:Observation`. The values of the LAI column populate the triples that describe the LAI values of the `Observation`, and the values of the columns `time` and `wkt` populate the triples that describe the time and location of the observations accordingly.

Given the mappings provided above, we can pose the the following GeoSPARQL query to retrieve the Leaf Area Index values and the geometries of areas

```

select distinct ?s ?g ?lai where {
?s lai:hasLai ?lai .
?s geo:asWKT ?g }

```

Notably, both translation steps are performed on-the-fly and only after a GeoSPARQL query is posed to the system.

This approach goes considerably beyond the previous version of Ontop-spatial that could only connect to an existing database with materialized tables, as well as the default, non-spatial version of Ontop and any other RDB2RDF system. OBDA systems traditionally connect to an existing database with materialized tables and access it before a query is fired in order to collect metadata, etc. The exact schema of the database tables is known beforehand. The approach that we propose in this paper is *schema-agnostic*: Ontop-spatial does not know the schema of the data as there is no database materialized. The virtual table is only created on-the-fly.

Ontop-spatial can be available as a GeoSPARQL endpoint, and thus can be used both by federation and interlinking engines. For example, one can use the interlinking tool Silk (<http://silkframework.org/>) to interlink Copernicus data that is accessible as RDF graphs using Ontop-spatial with linked data that is available using standard (Geo)SPARQL endpoints.

3. EVALUATION

We have evaluated Ontop-spatial by extending the benchmark Geographica with support for the evaluation of OBDA systems. Geographica (<http://geographica.di.uoa.gr>) was initially designed to evaluate the performance of geospatial RDF stores. We compared Ontop-spatial with the state-of-the-art geospatial RDF store Strabon (<http://strabon.di.uoa.gr>) [4]. Strabon has also been developed by our group and has been shown to be the most efficient geospatial RDF store available today [5]. Our evaluation showed that Ontop-spatial generally achieves *significantly better performance* than Strabon, often by orders of magnitude, when a large number of geospatial intermediate results are generated during the evaluation of a query. For example, Ontop-spatial is able to execute spatial selections and spatial joins against a 30 GB dataset that contains complex geometries (i.e., from points to polygons containing thousands of points) in less than a second. A summary of the experiments that we carried out is illustrated in Figure 2.

For the experiments, we used as set of spatial selection queries and a set of spatial joins. The queries in both sets contain a spatial filter. In spatial selections, one of the two arguments of the filter is a spatial constant, for example it can be a point or a linestring or a polygon. We experimented with different kinds of spatial constants with variant number of points per geometry in order to construct queries with various spatial selectivity. To construct spatial selections with low selectivity we used polygons that cover a large area so that most of the geometries of the dataset are contained in this area. We did the opposite to construct highly selective queries. We used the same approach for the spatial joins. The difference is that in spatial joins both arguments of the spatial filter are spatial variables, not constants. As shown in Figure 2, both systems have execution times at the same scale

in spatial selections, regardless of the spatial selectivity of the queries. The difference in the performance of the two systems increases in spatial selections where the geometries involved are more complex (i.e., polygons).

In spatial joins the difference in execution times between Ontop-spatial and Strabon increases even more, especially in queries with low selectivity where complex geometries are involved. This is because the SQL queries that are produced by Strabon contain larger number of joins in comparison with Ontop-spatial, because of the schema of the database that serves as a back-end of Strabon, and the fact that all geometries are stored in a separate table which is R-tree indexed. In Ontop-spatial, on the other hand, every dataset is imported into the database as a separate table. The geometries are stored in a separate column of this dataset and are indexed using R-tree. So when a spatial join involves geometries from two datasets, only these two tables will be involved in the query evaluation. This issue is very common in RDF stores, as triple stores by nature store information about triples, whereas the relational model is more compact. This is the reason why Strabon, that extends the RDBMS version of Sesame triple store, creates a database that is a lot larger than the one that Ontop-spatial uses.

Both the functionality and the performance achieved by Ontop-spatial make it the system of choice for making geospatial data available using linked data technologies.

Operation (geof:intersects)	Selectivity	Geometry types	Strabon	Ontop-spatial	Remarks
Spatial Selection	high	* (irrelevant)	100 msec	100 msec	
Spatial Selection	low	Point-Polygon	100 msec	100 msec	
Spatial Selection	low	Polygon-Polygon	500 msec	100-200 msec	
Spatial Join	high	Point-Polygon	< 1000 msec	< 1000 msec	
Spatial Join	high	Polygon-Polygon	100000 msec	100000 msec	
Spatial Join	low	Polygon-Polygon	>40 mins	10 mins	Sometimes the difference here is order(s) of magnitude

Fig. 2. Overview of evaluation results

4. CONCLUSIONS

In this paper we presented an extension of the OBDA paradigm for accessing vector and raster data, as well as data offered through DAP services. Our future work will concentrate on extending Ontop-spatial with the ability to query array databases. We also plan to improve the scalability of our system by extending it so that it complies with spatially-enabled cloud infrastructures.

5. REFERENCES

- [1] "Open Geospatial Consortium. OGC GeoSPARQL - A geographic query language for RDF data," OGC Candidate Implementation Standard, 2012.
- [2] K. Bereta and M. Koubarakis, "Ontop of geospatial databases," in *ISWC 2016*.
- [3] Andrej Andrejev, Dimitar Misev, Peter Baumann, and Tore Risch, "Spatio-temporal gridded data processing on the semantic web," in *DSDIS 2015*.
- [4] Kostis Kyzirakos, Manos Karpathiotakis, and Manolis Koubarakis, "Strabon: A Semantic Geospatial DBMS," in *ISWC*, Philippe Cudr-Mauroux and et al., Eds. 2012, vol. 7649 of *LNCS*, pp. 295–311, Springer.
- [5] G. Garbis, K. Kyzirakos, and M. Koubarakis, "Geographica: A Benchmark for Geospatial RDF Stores," *ISWC 2013*.

BIG DATA CHALLENGES IN GEOSS

Stefano Nativi¹, Joost van Bemmelen², Mattia Santoro¹, Guido Colangeli³

¹Italian National Research Council – Institute of Atmospheric Pollution Research, ²European Space Agency, ³RHEA Group

ABSTRACT

The Global Earth Observation System of Systems' (GEOSS) aim is to provide discovery, access and use functionality to heterogeneous Earth observations (data, services, models, information, etc.) from all over the globe for a broad range of users in both public and private sectors. This paper documents the challenges GEOSS is facing in the area of big data, i.e. the growing amounts of heterogeneous and dispersed Earth observations and provides an example to the Chinese Silk Road

Index Terms— GEOSS, User-centric Infrastructures, Big Data, GCI, GUI, Earth Observations

1. INTRODUCTION

In this era of growing number of satellites, aerial and in-situ data sources, the available resources and quantities of Earth observation (EO) data and information that can be used as input to environmental monitoring, developmental and economic activities are increasing every day. While the in 2002 launched ESA Envisat satellite, which had a payload of 10 sensors, provided around 350 TB/year, the currently flying ESA Sentinel-1, with a single SAR sensor, already provides some 1.8 TB/day and thus in the order of 650 TB/year.

These growing quantities of resources and data, their diversity, their distributed nature, and the fact that more and more data are becoming freely and openly available, also introduce new challenges and continue asking for new solutions [1]. Certainly not the least important challenge regards the ability for users to (1) connect to all data, find and quickly filter out the most suitable ones according to the work to be performed, and (2) to access and use discovered data, possibly in real-time.

Big Data is a radical shift rather than an incremental change for most of the existing digital infrastructures. The increasing availability of observations from sensors and models, coupled with the ever-growing computing power provided by new technologies including Cloud systems, enabled an entirely new approach to science based on data intensive scientific discovery. This has required innovative enabling technologies for the management, analytics,

delivery and presentation of large amount of data. The term big data encompasses all these aspects and it is one of the current major trends in data science and Information Technology. However the big data concept itself is elusive, bringing to many possible definitions according the different aspects the focus is put on [2].

2. USER-CENTRIC DATA INFRASTRUCTURES

In responding to such challenges, one should consider that there exist various types of users and usages of EO data, including in the public sector (e.g. data scientists/researchers and decision/policy makers), in the private sector (e.g. value adders and service providers), for education and the citizens. Some being more related to scientific use, others more to decision and/or policy making. Some interested in one domain of application, like disaster management, urban development, water resources management, food security, infrastructure management or sustainable agriculture just to name a few, others more in other domains. Some of them requiring access to the nearly final information like statistics or reports that are actually derived from the more 'raw' data, others that require to work 'closer' to the data themselves and would need as well access to models and/or algorithms and computing resources where they can perform remotely their analysis and/or run simulations on data from different heterogeneous resources, since downloading all of them and processing them locally is no longer an option for most of them. So most users need the possibility to search, access and use the data remotely, and for this they need to have access to the right instruments – *tools, APIs, services, etc.* – that allow them to do so.

3. THE GEOSS EXPERIENCE

The Global Earth Observation System of Systems (GEOSS) [4] is a global and flexible network of content providers allowing decision makers to access an extraordinary range of data and information at their desk. The GEOSS Common Infrastructure (GCI) has implemented a software ecosystem by applying the upstream-midstream-downstream pattern to connect the Users to the vast supply system: the GEO-DAB (Discovery and Access Broker) and the GEOSS Portal play an important role to this end. For instance, Figure 1 depicts GEOSS discovery and access capabilities in a selected region that covers Asia, Africa, and Europe addressing the

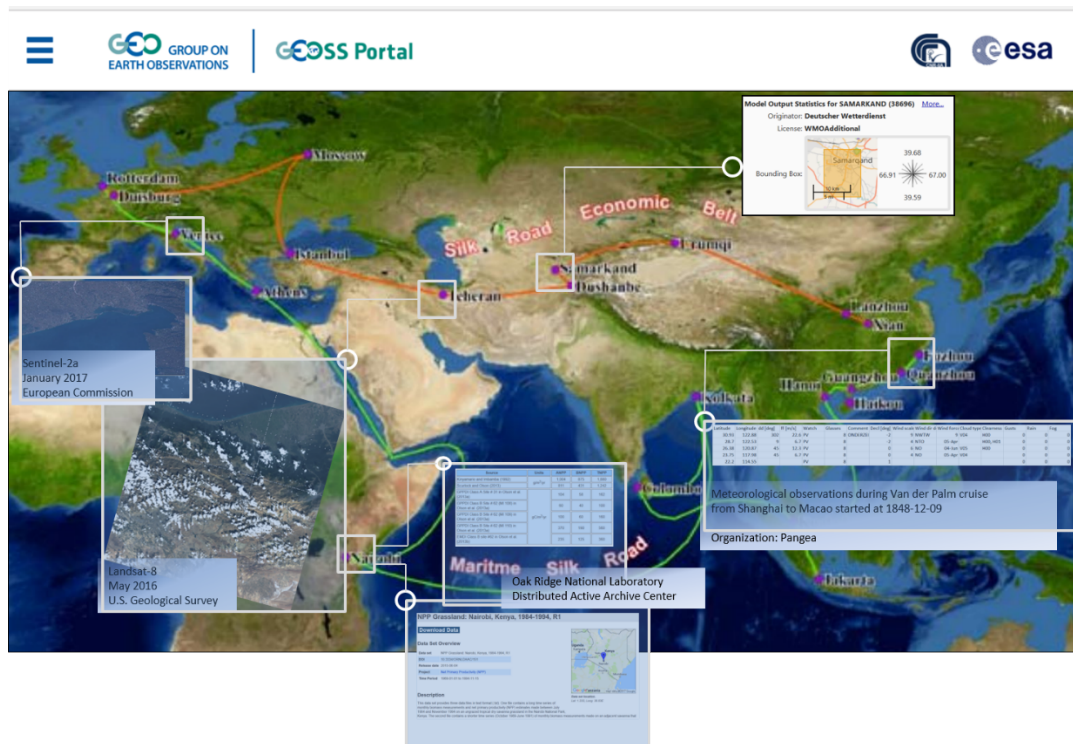


Figure 1 - GEOSS capabilities showing data records from different heterogeneous data resources that could be of interest to the DBAR Community

needs of the Community working for the “Digital Belt and Road” (DBAR) initiative [3].

4. GEOSS AND THE BIG DATA CHALLENGES

The impact of the big data dimensionalities (i.e. the ‘V’ axes) on the GCI is summarized by the following table, describing the solutions and strategies adopted – particularly, for the GEO-DAB and the GEOSS Portal – in relation to the challenges of Discovery, Access and Use of Earth Observations.

With respect to the Big Data challenges, the role of GEO-DAB/GEOSS Portal platform in the GCI is to enable a seamless and homogenous exploitation of resources contributed by the GEOSS data providers. The particular focus of the GEO-DAB is to address the challenges in terms of interconnection with the GEOSS data providers’ systems, while the GEOSS Portal is in charge of presenting results to users in the way which is most appropriate taking into account users’ needs and available results.

Big Data challenges faced by GEOSS		GCI Solutions used to address the challenges
VOLUME	<p>Discovery Challenges</p> <p>high number of catalogs, inventory, listing services to be discovered;</p> <p>Large number of metadata records;</p> <p>Large number of Users’ discovery requests</p>	<p>Reduce the number of matching results, by supporting GEOSS Views. A view is a subset of the whole GEOSS resources defined by applying, via the DAB, a set of Discovery and/or Access clauses.</p> <p>Design and apply a ranking metrics and related paging strategy for presenting long results sets in a more useable and faster manner.</p> <p>Support distributed queries, along with harvesting approach, to reduce the number of large metadata records to be stored and managed by the DAB.</p> <p>Use of load balancing and auto-scaling clusters to support large number of queries.</p>
	<p>Access/Use Challenges</p> <p>high number of data services to be accessed;</p> <p>large amount of datasets;</p> <p>big data volume;</p>	<p>Use of server-side transformation functionalities to limit downloaded data to that which will be effectively leveraged by GEOSS Users.</p> <p>Supplement missing transformation functionalities (not supported by data servers) to limit data downloaded and processed.</p> <p>Support data caching and map tiling through NoSQL technology.</p>

	Large number of Users' access requests	Use of load balancing and auto-scaling clusters to support large number of access requests.
VARIETY	<p>Discovery Challenges</p> <p>Support of highly heterogeneous metadata models and discovery service interfaces;</p> <p>Publication of the set of metadata models and discovery interfaces implemented by GEOSS Users' applications;</p> <p>Long-term data access sustainability in a multidisciplinary environment</p>	<p>Introduction of a brokering tier dedicated to mediation of service interfaces and metadata models harmonization in a transparent way for both Users and data providers. This applies the following interoperability principles:</p> <ul style="list-style-type: none"> • Adopt a solution preserving existing Users' applications and data systems autonomy. • Supplement but not supplant disciplinary infrastructures mandates by interconnecting and mediating them. • Minimize the technological and organizational barriers for both Users and data providers to be interconnected. <p>Design and implementation of a brokering semantic and metadata model used to: (i) harmonize and integrate the heterogeneous metadata models brokered by GEOSS; (ii) expose the metadata views well-supported by GEOSS Users.</p> <p>Extensible architecture of brokering to support new service interfaces and metadata models.</p> <p>GEOSS Portal customization for SBA/thematic-specific requirements and data-types (e.g. ad-hoc filters and result visualizations).</p>
	<p>Access/Use Challenges</p> <p>Support of highly heterogeneous data models, encoding formats, and access service interfaces;</p> <p>Publication of the set of data models, encoding format, and access interfaces implemented by GEOSS Users' applications;</p> <p>Long-term data access sustainability in a multidisciplinary environment</p>	<p>Introduction of a brokering tier dedicated to mediation of access service interfaces and data formats harmonization in a transparent way for both Users and data providers. This applies the following interoperability principles:</p> <ul style="list-style-type: none"> • Adopt a solution preserving existing Users' applications and data systems autonomy. • Supplement but not supplant disciplinary infrastructures mandates by interconnecting and mediating them. • Minimize the technological and organizational barriers for both Users and data providers to be interconnected. <p>Design and implementation of a brokering data model used to: (i) harmonize and integrate the heterogeneous data formats brokered by GEOSS; (ii) expose the data formats well-supported by GEOSS Users.</p> <p>Extensible architecture of brokering to support new access service interfaces and data formats.</p> <p>Transformations facilitating re-use, i.e. making data available with a Common Grid Environment (CRS, resolution, extent, format).</p>
VELOCITY	<p>Discovery Challenges</p> <p>To manage the increasing rate at which metadata flows;</p> <p>Fast metadata processing to satisfy Users' needs</p>	<p>Operational data store that periodically extracts, integrates and re-organizes (harvests) brokered metadata records for operational inquire and ranking generation.</p> <p>Caches that provide instant access to the results of distributed queries while buffering data provider systems from additional load and performance degradation.</p> <p>Design of the DAB architecture that balances metadata latencies with GEOSS Users' requirements, avoiding to assume that all data must be near-real time.</p> <p>Use of No-SQL document store with selection of indexed elements for flexible ranking.</p> <p>Incremental harvesting strategy allowing to limit the number of resources handled during each re-harvesting.</p> <p>Live query distribution combined with caching of results allowing to speed up the retrieval of non-harvested records.</p> <p>Load balancing to route incoming requests to machines with lowest workload.</p> <p>Use of auto-scaling clusters to increase computing capacity in response of rapid workload growth.</p>
	<p>Access/Use Challenges</p> <p>To manage the increasing rate at which data flows;</p> <p>Fast data processing to satisfy Users' needs</p>	<p>Operational data store that periodically generates and stores preview tiled maps of brokered data for operational data preview.</p> <p>Caches that provide instant access to the results of previous access requests while buffering data provider systems from additional load and performance degradation.</p> <p>Supplementing missing transformations allows limiting the local processing time.</p> <p>Store and retrieve caches and preview tiles through NoSQL key-value store.</p> <p>Load balancing and auto-scaling (see <i>Discovery</i> row above).</p> <p>For extremely large processing requests Users are allowed to opt for an asynchronous version of the access functionality.</p>

VERACITY and VALUE	<p>Challenges</p> <p>Reduce the “information noise”;</p> <p>Retrieved data comparison;</p> <p>Data trustiness for GEOSS decision makers;</p> <p>Effective data re-use;</p> <p>Data meaningfulness for User requests;</p> <p>Data accuracy for intended use</p>	<p>The brokering data model includes a specific multi-disciplinary quality extension.</p> <p>Implementation of a flexible ranking metrics including quality of service and metadata completeness as valuable indexes.</p> <p>The brokering metadata model supports a harmonized presentation of retrieved metadata facilitating their comparison.</p> <p>Use of GEOSS Essential Variables as an additional parameter for improving the existing ranking metrics.</p> <p>The prototyped “fit-for-purpose” and Users’ feedback extensions aim to provide Users with quality-aware results.</p> <p>Definition of GEOSS Data management principles including quality-related aspects.</p> <p>Display of relevant information in a structured way by GEOSS Portal at user request only.</p>
VISUALIZATION	<p>Challenges</p> <p>Visualization speed;</p> <p>Contextualized visualization</p>	<p>Use of latest web-technologies considering powerful scripting tools, introduction of new tools (plug & play), multi-modality, social and collaboration enhancements.</p> <p>Support of reusable GEOSS Portal portlets for integration in external community applications.</p> <p>Customization of GEOSS Portal for SBA-specific requirements.</p> <p>Support Community Portals and Applications by publishing DAB APIs for client development.</p> <p>Support the following visualization strategy: (1) provide an overview (trying to keep that simple and show important elements), (2) allow zoom and filter unnecessary clutter, (3) provide more details if requested by Users.</p> <p>Use of Common Grid Environment (CRS, resolution, extent, format) functionality to generate previews of accessible data.</p> <p>Use No-SQL key-value databases to store and retrieve previews of accessible data, combined with a proper key generation.</p> <p>Provide fast previews by generating preview tiles in batch.</p>

5. CONCLUSION

Challenges to discover, access and/or use Earth observations are not new – different examples exist of usage of emerging technologies to deal with such challenges. As well GEOSS, via the implementation of the GEOSS Common Infrastructure and in particular with the Discovery and Access Broker and the GEOSS Portal, is dealing with these challenges, in particular, knowing that GEOSS has over one hundred million of resources brokered (and this number is growing every day), and many different interested parties with different interests and backgrounds (scientists, value adders, decision makers, policy makers, citizens, ...) that need these resources to be discoverable and accessible. Different approaches and solutions are adopted as reported in this paper. Future evolutions and enhancements of the GCI and GEOSS at large are however still needed, not only because of the growing amounts of heterogeneous data, but for sure as well because of the different users and new usages across time, space and the need to consider all the available data, information, knowledge and wisdom.

6. REFERENCES

- [1] J. van Bemmelen, L. Fusco, V. Guidetti, 2005, Access to distributed Earth Science Data Supported by Emerging Technologies, The 19th international conference EnviroInfo 2005 – Informatics for Environmental Protection, Brno, Czech Republic, September 7-9, (2005) ISBN: 80-210-3780-6.
- [2] S Nativi, P Mazzetti, M Santoro, F Papeschi, M Craglia, 2015, “Big data challenges in building the global earth observation system of systems”. Environmental Modelling & Software, Vol. 68, pp. 1-26.
- [3] Guoqing Li, Bopha Silap, Nativi Stefano, Bemmelen Joost van, Santoro Mattia, Colangeli Guido, Zhe Xu, Zhifeng Guo, Romero Laia, Martinez Bernat, Jing Zhao, Chuanzhao Tian, 2017, Well Using of Big EO Data to Support Geo-earth Cognition of the Belt and Road, Bulletin of Chinese Academy of Sciences, 2017, 32(Z1): 10-17.
- [4] Group on Earth Observation, 2017, About GEOSS, available at: <https://www.earthobservations.org/geoss.php>.

COPERNICUS AND AIS DATA FUSION AND INFORMATION MANAGEMENT FOR MARITIME TASKS – PRELIMINARY RESULTS

*José Manuel Delgado Blasco¹, Claudio Manganiello², Pier Giorgio Marchetti³,
Massimo Marrazzo², Mauro Arcorace¹*

¹ Progressive Systems srl., ² Italian Coast Guard, ³ European Space Agency

ABSTRACT

This paper describes the preliminary results and proposed solutions from a study on Copernicus satellite data fusion with maritime navigation data. The context of the proposed study is the use of Copernicus data for the fulfillment of maritime tasks by the Italian Coast Guard. We will discuss the issues related to data volume, velocity and variety, as well as proposed solutions addressing approaches for data reduction, information management and visualization. The paper focuses on the use of open source tools. The specific use case for this study is the one of maritime traffic monitoring.

Index Terms— data fusion, Earth observation, maritime traffic monitoring, Copernicus Programme, Sentinel-1, Sentinel-2, Automatic Identification System

1. CONTEXT, OBJECTIVES AND DATA SOURCES

The tasks of the Italian Coast Guard include: search and rescue, maritime law enforcement, maritime and coastal protection of marine resources, maritime safety and security, fisheries protection and regulation. The extension of the Italian marine coasts is more than 8000 km, and the nominal search and rescue region is about 500000 km² [1]. The area can more than double - i.e. get to 1130000 km² - in case of management and/or coordination of search and rescue operations in the geographic areas of Sicily Channel, South Ionian and Libyan sea. The area, which falls within the “Central Mediterranean migration route” [2], is highly demanding both in terms of operations as well as of international coordination of operations.

The maritime traffic in the Sicily Channel is very busy – 100-150 vessels per day - as this is the preferred route for cargo ships and tankers between the Suez Channel and Gibraltar. For the purpose of this study, an area of interest around the island of Pantelleria was defined for the processing of Sentinel-1 and Sentinel-2 products and the prototyping of near real time data fusion with Automatic Identification System (AIS) data [3], [4] and other vessel location information sources, for simplicity in the following collectively referred to as “AIS”.

Within the Copernicus Programme [5], the data acquired by the Sentinel missions is offered through an open and free data policy [6]. Furthermore, the Sentinel missions have the operational objective to provide global, timely and easily accessible information in application domains such as land, marine, atmosphere, emergency response, climate change and security.

The Sentinel missions offer a direct response to Earth observation data needs for maritime surveillance [7], [8] of institutions like the Italian Coast Guard. As a matter of fact, Sentinel (mission) data provides a complementary source of data and information which can be fused and integrated with own data sources (e.g. the AIS data stream) to accomplish effectively and efficiently institutional tasks.

Sentinel-1 synthetic aperture radar (SAR) mission observation scenario did foresee from its design [9], the systematic coverage of land, Arctic and Antarctic, and of the Mediterranean waters, as well as of other European Union relevant shipping routes. The availability of two satellite units, establishing a constellation allows six-day exact repeat and conflict-free operations based on two main operational modes. Being the Mediterranean a sea between the two land areas of Europe and Africa, the operating mode of the radar remains the most suitable for land observation, i.e., the Interferometric Wide-swath mode [9], [10] even when the Sentinel-1 satellites are overpassing the Mediterranean waters. Still the Sentinel-1 constellation offers the capability to detect ships starting from the 20-25m length or even less in “ideal” sea conditions.

The objectives of the study were:

- To perform data fusion between the synthetic aperture radar Sentinel-1 SAR data and the AIS data with the purpose to analyze and eventually quantify: a. non-cooperative ships transit in the study area; b. small vessels (fishing, sailing or motor boats, other) located in the study area.
- To assess the usability of Sentinel-2, optical satellite data, for the same purposes.
- To define and test the methods for data reduction and data fusion, needed to satisfy the above objectives within an operational context.

An additional project constraint was to use only open source software for the Sentinel data processing. Two open source tools were selected:

- the Sentinel Application Platform (SNAP), a.k.a. Sentinel Toolbox [11] for the Sentinel-1 and Sentinel-2 data processing and visual analysis including initial inspection of AIS data and
- the QGIS [12] geographic information system, whose vector geo-processing tool has been used to validate the python scripts created for the data fusion step.

The data streams addressed in the study were AIS real time data, and the Sentinel-1 and Sentinel-2 data acquisitions over the study area.

The real time AIS binary data stream from the Pantelleria station study area is a binary stream of about 5 Kbit/s, who needs to be decoded and stored on the prototype server, until the data fusion step is performed. Due to the different obligations to which different type of vessels are subject, the temporization with which the AIS data stream is transmitted by the vessel (and received when the ship is in visibility of the receiving station), vary from few seconds to several minutes [3],[4]. The time axis is therefore characterized by a discrete non uniform availability of AIS data with a granularity of the second, to which needs to be superimposed the discrete availability of Copernicus data, which - at the best - considering the two by two satellite constellations of Sentinel-1 and Sentinel-2 means availability of data every day and half.

To be able to exploit all the time the best resolution possible, the Sentinel-1 mission data needs to be downloaded and stored temporarily both in the Single Look Complex (SLC) format and in the Ground Range Detected (GRD) [13]. As a matter of fact, the processing SLC data allows to exploit at maximum the resolution of the synthetic aperture radar (5 by 20 meters) at the expense of a longer processing time and a workflow encompassing de-bursting, calibration, object detection [14] and terrain correction.

Using the GRD data offers a shorter object detection processing time (about one third of the one needed by SLC) at a nominal resolution of about 20 by 20 meters. Further experimentation has been conducted running a simplified threshold based formula - named the 'Close' formula (defined by Claudio Manganiello and Jose' Manuel Delgado Blasco) - which can be used for quick object detections in "ideal" sea and wind conditions. For the purpose of this study we considered "ideal" condition a wind force ≤ 2 Beaufort [15] during 12 hours.

The Sentinel-2 data stream has been limited to products with a cloud coverage of less than 20%. The Sentinel-2 products

are then processed to produce a mask computed by a modified version of the second Normalized Difference Water Index (NDWI2) [16].

Additional study work has been dedicated to the data processing and reduction as well as to the management of the flow of information. The data flow needs to be automatized for the recurrent pre-processing and processing activities. It shall minimize the amount of data which is permanently stored in the prototype environment, which shall be limited to near real time information extraction or data re-processing.

In terms of product requirements about 160 Sentinel-1 products and about 100 Sentinel 2 (A unit, 20% cloud cover) are required to cover the Italian search and rescue area over the nominal repeat cycle of 12 days.

The results based on running the prototype over the first three quarters of 2017 show that a temporal moving window of seven days is enough to support the main operational tasks. Therefore, an architecture based on a master server (equipped with 16 CPUs, 32Gbyte RAM and 16Tbytes HDD/SDD) collecting both the AIS data and Copernicus data on shared disk units, complemented by 5 similar, servers dedicated to complementary geographic areas - coupled with an efficient data cleaning process - could be sufficient to cover the processing and information extraction needs for the area of interest of the Italian Coast Guard.

A conceptual separation can be performed between the automated data-preprocessing and information extraction steps and the visual analysis steps which need to be performed by an expert. The prototype environment used within this study took into account the research work in [17] in particular concerning the sequence of human-system interactions, the timing of Copernicus data availability, as well as the data representation and best practices adopted within the operational environment used by the experts who could perform the visual analysis for the prototype validation.

2. DATA PROCESSING AND FUSION

In terms of CPU time, memory usage and processing time, the most demanding data processing workflow is the one related to the "object detection" (ship detection) both on SLC and GRD Sentinel-1 products. In SNAP the object detection is implemented using the classical Constant False Alarm Rate (CFAR) algorithm [14]. The detector compares pixels from the target cell within a sliding window to a threshold depending on the statistics of the backscatter of the surrounding area. On the basis of the extensive data analysis performed in [18] and [19], we decided to use the classical adaptive threshold algorithm CFAR available on SNAP, assuming the Gaussian distribution for the clutter

both for the SLC - on demand - and GRD – systematic - data processing. The data analysis performed during our work could confirm the conclusion of [18] that VH (cross) polarization achieves better target to clutter ratios than VV. For the sake of completeness, we have to report that large ships may generate ghost signals due to the so-called azimuth ambiguity [20], [7]. The ghost signals are not automatically removed by current prototype processing chain.

Furthermore, it has to be noted that we could observe that the CFAR detector on Sentinel-1 data may misbehave due to the presence of heavy rain cells and gust fronts, which were observed near the test area in winter 2017. Figure 4 in [21] shows a 3D representation of the rain and wind behaviors which well models the observed data features.

Over the study area, the processing workflow performs data reduction of a factor about 10^6 : from about 1.5 GB for each compressed GRD product to few KB of comma separated values representing the detected objects. The workflow is executed systematically (at pre-defined time intervals) by a job which invokes the SNAP command line interface towards the Graph Processing Tool [22]. The “object detections” in SNAP terminology, i.e. the targets are characterized by the coordinates and detected ship length and width. The estimation of the latter two parameters could not be used in the data fusion step, due to the mismatch with actual vessel characteristics. Future work should benchmark and possibly use the open source software developed in [7],[8] and focus on the 3D representation of detected (non-cooperative) vessels from the SAR product VH intensity data.

The data fusion step has the purpose of eventually triggering further analysis by an operator, henceforth – always with the objective to perform a data reduction – the AIS stream analysis within the test station is performed considering only the AIS data acquired in the 5 minutes around the Sentinel-1 satellite pass. The AIS data may then be directly visualized in the SNAP viewer as “AIS pins” together with the SAR “object detections” for visual analysis. The analysis shall take into account that the position of the detected “objects” is affected by a displacement w.r.t the actual coordinates caused by the SAR Doppler shift which is in turn depending on the relative direction and speed of the “object” respect to the satellite orbit.

A subsequent step evaluates the SAR ship detections that match the AIS data by using python scripts exploiting geospatial functions. By defining a “buffer” around the *AIS vessel position* and the *SAR detected object* they identify the intersected and non-intersected objects, by fusing the information coming from the two data streams i.e. Sentinel-1 SAR and AIS. The scripts allow to test the most suitable buffer size. In our prototype it has been set to 0.01degrees,

which at the latitude of the study means about 0.5 nautical miles. The buffer size shall take into account both the inaccuracy in the position detected by the SAR discussed above, as well as the fact that the AIS position information being discrete in time, may not be taken at the same time of the SAR observation. The outcome of this intersecting operation are two datasets containing: i) the set of intersected (*AIS vessel position* and the *SAR detected object*) objects, called *cooperative targets* and; ii) the set of *SAR detected object* for which no *AIS vessel position* data is available, called *non-cooperative targets*.

Later, the resulting datasets as well as the processed SAR image can be loaded in a geospatial information system, such as QGIS [12], allowing the experts to perform their analysis on the “cooperative” and “non-cooperative” targets.

3. THRESHOLD BASED DATA ANALYSIS

The objective of this task was to identify – given the “ideal” sea conditions defined above - a simplified threshold based SAR data analysis approach which could be used for a quick analysis on near real time data in operational context which requires very fast identification of small vessels. To fulfil this purpose some tests have been run using the so-called Close’ interval. The Close’s interval defines a SAR calibrated backscatter σ_{0dB} interval for both VV and VH polarization which has the purpose to level the backscatter and characterize only objects located in the middle of the sea. This interval has been identified empirically for “ideal” sea conditions and it is defined as follows:

$$Close' interval = \sigma_{0dB}^{VV} > -14 \text{ AND } \sigma_{0dB}^{VH} > -20 \quad (Eq.1)$$

Using the Close’s interval, we create a binary mask for each SAR image, as follows, named Close’s formula:

$$Mask(x) = \begin{cases} 1, & \forall x \in Close' interval \\ 0, & \forall x \notin Close' interval \end{cases} \quad (Eq.2)$$

By using this procedure, we manage to reduce each SAR image into a single binary mask. In a later stage, each mask is converted into comma separated files with the geographical information of the objects detected, reducing even more the output data.

The Sentinel-2 products are processed to produce a mask computed by a modified version of the second Normalized Difference Water Index (NDWI2), the threshold value is close to 0.2, however heavily depending on many factors i.e. clouds [16].

4. ACKNOWLEDGEMENTS

The authors want to thank Antonio Vollero from the Italian Coast Guard for the support provided in the use of AIS data

in the study area, and Pierre Potin, Nuno Miranda and Jolyon Martin from ESA for the support in the use of Copernicus data.

The authors want to thank as well, the whole Research and Service Support team [23], in particular, Giovanni Sabatino and Błażej Fitrzyk, for the continuous support during the prototype development, the development of key geospatial analysis modules and the assistance provided in setting up the cloud-based virtual environment used for the prototyping and the high volume data management performed. The team provided as well expert support on SAR and optical data exploitation, and ensured the continuous running of the prototype environment.

5. REFERENCES

- [1] “MULTILATERAL International Convention on maritime search and rescue”, [UN Treaty No. 23489](#), 1979
- [2] “Migration on the Central Mediterranean route Managing flows, saving lives”, [JOIN\(2017\) 4 final](#), European Commission, Brussels, 25.1.2017
- [3] “RECOMMENDATION ON PERFORMANCE STANDARDS FOR AN UNIVERSAL SHIPBORNE AUTOMATIC IDENTIFICATION SYSTEM (AIS)”, IMO, RESOLUTION MSC.74(69), ADOPTION OF NEW AND AMENDED PERFORMANCE STANDARDS, 1998
- [4] Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band [.ITU-R, Recommendation ITU-R M.1371-5 \(02/2014\)](#), 2014
- [5] “REGULATION (EU) No 377/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 3 April 2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010”, [Official Journal of the European Union](#), 2014
- [6] J. Aschbacher, M. P. Milagro-Pérez, “The European Earth monitoring (GMES) programme: Status and perspectives”, *Remote Sensing of Environment* 120 (2012), 3–8, ELSEVIER, 2012
- [7] C. Santamaria, M. Stasolla, et al., Sentinel-1 Maritime Surveillance - Testing and Experiences with Long-term Monitoring, JRC Science and Policy Reports, *Publications Office of the European Union*, 2015
- [8] C. Santamaria, M. Stasolla, et al., Mass Processing of Sentinel-1 Images for Maritime Surveillance, *Remote Sensing*, 2017, 9, 678; doi:10.3390/rs9070678 , 2017
- [9] R. Torres, P. Snoeij et al., “GMES Sentinel-1 mission”, *Remote Sensing of Environment* 120 (2012), 9–24, ELSEVIER, 2012
- [10] P. Potin, P. Bargellini et al. “SENTINEL-1 MISSION OPERATIONS CONCEPT”, *Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, 2012
- [11] Y.-L. Desnos et al. “SCIENTIFIC EXPLOITATION OF SENTINEL-1 WITHIN ESA’S SEOM PROGRAMME ELEMENT”, *Proceedings of Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, 2016
- [12] T. Sutton et al. “Documentation for QGIS 2.18”, <http://docs.qgis.org/2.18/en/docs/index.html> , accessed 15 May 2017
- [13] AA.VV. “Sentinel-1 User Handbook “, GMES-S1OP-EOPG-TN-13-0001, ESA, 2013
- [14] D.J. Crisp, “The State-of-the-Art in Ship Detection in Synthetic Aperture Radar Imagery”, *DSTO Information Sciences Laboratory*, DSTO-RR-0272, 2004-05, 2004
- [15] “Manual on Marine Meteorological Services – Volume I Global Aspects”, *World Meteorological Organization*, [WMO-No. 558](#), 2012
- [16] S. K. McFeeters, “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”, Volume 17, Issue 7, 1996
- [17] M. Riveiro, G. Falkman, “Supporting the analytical reasoning process in maritime anomaly detection: evaluation and experimental design”, *Proceedings of the 14th International Conference Information Visualisation*, 171-178, 2010
- [18] R. Pelich, et al., “Performance evaluation of Sentinel-1 data in SAR ship detection”, *Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*. IEEE, 2015
- [19] R. Pelich et al., “AIS-Based Evaluation of Target Detectors and SAR Sensors Characteristics for Maritime Surveillance”, *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, Vol. 8, No. 8, August 2015 3892-3901, 2015
- [20] J.C. Curlander, R.N. McDonough, Synthetic Aperture Radar Systems and Signal Processing, *John Wiley and Sons, Inc.*, 1991
- [21] W. Alpers et al. “Rain footprints on C-band synthetic aperture radar images of the ocean – Revisited”, *Remote Sensing of Environment* 187, 169–185, ELSEVIER, 2016
- [22] N. Fomferra et al. “SNAP Home, Developer Guide”, <https://senbox.atlassian.net/wiki/display/SNAP/SNAP+Home>, accessed on 18.05.2017
- [23] P.G. Marchetti, G. Rivolta et al., “A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support.”, *IEEE Geoscience and Remote Sensing Society Newsletter*, Vol. 162, 10-18, 2012

LESSONS LEARNED OVER THE PAST THREE YEARS, ON THE BIG DATA USAGE FOR PROCESSING GAIA DATA IN CNES

Frédéric Pailler¹, Laurence Chaoul¹, François Riclet¹, Chantal Panem¹,
¹CNES, 18, avenue Edouard Belin 31401 TOULOUSE CEDEX 9, France

ABSTRACT

Gaia is an astronomy mission of ESA, based on a satellite launched on 19th December 2013. Its main goal is to map more than one billion stars and sky objects. The scientific data processing is delegated to the Data Processing and Analysis Consortium (DPAC), and relies on six Data Processing Centres (DPC) distributed all around Europe. The CNES Data Processing Center (DPCC) is one of the main DPCs and has set up an architecture based on the Hadoop technology. After more than 3 years of usage, this paper will present the main lessons learned using Big Data technologies for Gaia scientific processing.

Index Terms— Processing Centers, Big Data, Hadoop, MapReduce, Spark

1. OVERALL ORGANISATION AND DATA FLOW

The Gaia satellite is located at Lagrange point L2, 1.5 million kilometers away from earth, in the opposite direction of the sun. The satellite scans the sky with combined rotations, and records objects detected by its two on-board telescopes.

Every day, it sends an average of 30 GB of telemetry to earth, received by three ESA ground stations. These data are transferred to the DPCE at ESAC (Madrid). DPCE hosts the Science Operations Centre (SOC) and runs the two core systems managing the telemetry: Initial Data Treatment (IDT) and First Look (FL). Their results are daily distributed to the other DPCs, which process them to monitor the payload health and to raise science alerts. They are called **daily chains**.

But the main goal of Gaia is to produce a catalog. Several versions are planned, each being the result of a processing **cycle** during 1 or 2 years. Cycle 1 produced the first catalog on 14th of September 2016 [DR1]. Cycle 2 is on-going, with a DR2 release planned on April 2018. A total of 4 versions are planned to the final Gaia catalog [DR Scenario].

Each catalog is the result of reprocessing **all** the data from the beginning of the mission (growing data volumes), with improved software (growing processing demand). They are called **cyclic chains**. In each cycle, each DPC produces its own results, computed by the cyclic chains, and sends them to DPCE where they are integrated into the Gaia Main DataBase (MDB), and finally used to create the catalog.

The DPCC is in charge of:

- two daily chains (spectrometer monitoring and asteroid detection),
- several cyclic chains computing radial velocities and astrophysical parameters, or detecting and classifying outlier objects (galaxies, quasars, solar system objects, non-single stars, etc.).

The daily and cyclic chains have fundamental differences:

- Daily chains run every day and process a constant volume of data. The processing time and the volume of results are predictable.
- Cyclic chains process a growing amount of data from one cycle to the next, with significant improvement of the software: the volume of data produced and the required amount of processing can only be assessed based on estimations.

The chains are built upon software code written by astronomers.

2. THE DPCC MAIN CHALLENGES

Gaia will observe 80 times more than 1 billion stars. The ground data processing has therefore to face several challenges:

- A huge **number** of elements to handle dozens of tables containing up to 80 billion rows;
- A complex processing with different **calendar constraints**: short delays for daily processing, vs final deadline for cycle;
- Huge **volume** of data: 3PB of results are foreseen at the end of the mission (not including intermediate data);
- **Resource sharing** between several processing chains to cope with the calendar constraints.

3. THE DPCC ARCHITECTURE

3.1. Data exchanges at DPCC

Data exchanges at DPCC are asymmetric: DPCC receives more data than it sends.

For daily processing, DPCC receives all the DPCE core systems results (IDT and FL). The average volume is 200 GB per day. The results of the daily processing are published to a dedicated web server (Gaiaweb) where the

scientists download what they need to perform the payload monitoring.

For cycle processing, DPCC receives data less often but in batches of typically several tens of terabytes in a few days. The results of the cyclic chains, much smaller in volume, are transferred back to DPCE to be stored in the MDB.

DPCC also receives all MDB data, which are saved in a backup MDB.

All these received data are temporarily stored into intermediate file system (Figure 1). Then the data are checked, sorted and inserted into relevant DPCC internal systems, using several software tools.

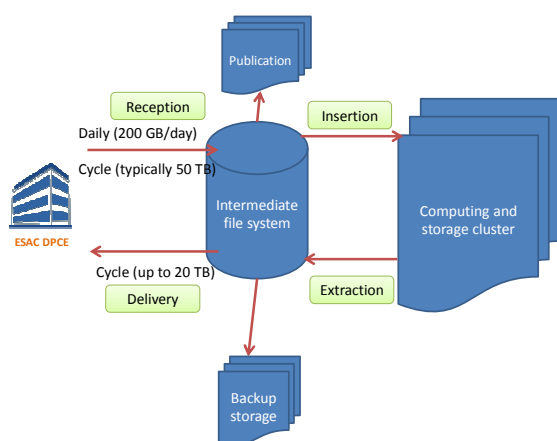


Figure 1: overview of the DPCC data flows.

3.2. An architecture based on Hadoop

To face the different challenges described in chapter 2, the CNES selected in 2010 the Apache Hadoop solution, with the Cloudera distribution.

Hadoop offers a highly distributed framework, with a powerful distributed file system (HDFS). It can manage the resources allocation to each processing chains. This ensures the availability of resources and guarantees the performances needed to be ready for the Gaia Data Releases schedule.

Furthermore, Hadoop provides the scalability allowing incremental hardware upgrades, in order to follow the growing needs in terms of volume and processing power over the 5 years of the mission. Hadoop can also manage heterogeneous hardware, so different hardware providers can be chosen at each upgrade.

The map/reduce programming model was implemented for most of the processing chains, through Cascading.

3.3. Hardware overview

The DPCC cluster is composed of high density nodes. Each node provides both computing power and storage space. The DPCC design is entirely scalable: the cluster performances are proportional to the nodes number.

The current operational platform contains 172 calculus nodes (3500 cores), with a total hard disk capacity of 2.6 PB and 19 TB of memory. The foreseen final DPCC operational cluster will contain 5800 cores/4.5 PB, stored in a set of 4 racks.

4. LESSONS LEARNED

After more than 3 years using Hadoop at DPCC, the main lessons learned are described below.

4.1. Lessons learned about data management

4.1.1. Intermediate file system

In the initial architecture the intermediate file system was implemented with GlusterFS, distributed on the different nodes of the cluster. Its main advantages were to be scaled to the cluster, and to grow when the cluster was upgraded. This system was running well with a few nodes, but started to show its limits when the number of nodes reached 20: some data loss appeared even with a replication 2 factor. Moreover, using a file system distributed on all the nodes was not efficient: transfers from the intermediate file system to HDFS were very slow and using a lot of network throughput.

The solution was to modify the intermediate file system the following way (in summer 2017):

- GlusterFS was replaced by MooseFS,
- Instead of being distributed into all the nodes of the cluster, the file system was installed into a set of dedicated servers.

This solution offers so far the required robustness and performance.

4.1.2. Data replication in Hadoop

The way to store the data in HDFS is an important point. Hadoop manages data replication in HDFS. The Hadoop recommendation is to use a replication factor 3 to ensure no data loss. But this requires a storage capacity three times higher than the data volume. According to the data size estimation at the end of the mission, DPCC decided to adapt the replication factor according to the criticality of the data: 3 for the input and output data and only 2 for intermediate data which can be recomputed if needed.

4.1.3. The blocksize puzzle

Gaia data are stored in HDFS as files with a given blocksize. This blocksize is an important parameter because the number of maps (of the map/reduce model) is defined by Hadoop as the number of blocks of the input data (1 map per block).

Moreover, in Hadoop V2 the number of maps of a job is limited to 100.000. Hence it is important to carefully configure the blocksize of input data to ensure this limit will not be reached.

And finally, when a job is split into too many maps, each map will have to perform very small computations, leading to a significant overhead of the map creation with respect to what it computes.

On the other hand, a too small number of maps underuses the resources, and globally takes longer.

The size and the number of the Gaia data records (scientific data) are very heterogeneous, and the computations are chained. Both can lead to non-optimal number of blocks, and global efficiency loss.

So a specific blocksize must be applied to each step of the chains. This tricky configuration is defined during the first validation tests of the chains.

Moreover, a limit to the number of reduces has been set to avoid output data fragmentation. Combined to a consistent blocksize, the number of maps of the next step of the chain can be kept under control.

4.1.4. Data queries in HDFS

Hadoop is a NoSQL database. Hence depending on the data and on which computation is done, the efficiency can be dramatically different.

The ideal situation is when each data record can be processed independently to the others. For example, processing one billion stars individually to compute their magnitudes will be very efficient. Each core will process a subset of the stars and does its job without knowing what the others do.

But when queries or filters are applied, the efficiency drops. For example selecting a subset of the stars located in a given area of the sky, or stars observed inside a given a time window, are inefficient cases.

A solution would be to define metadata in HDFS to increase efficiency, but this was not implemented at DPCC so far.

4.2. Lessons learned about data processing

4.2.1. Fair scheduling

In Hadoop different queues can be configured to limit the resources used by the chains. This prevents one chain from consuming all the cores or all the memory, and provides a way to share the resources of the cluster.

The Fair Scheduler has been set up at DPCC: a hierarchy of queues is configured with, for each, limits in number of cores and amount of memory. The Resource Manager distributes fairly the resources to all the chains running at a given time.

The queue parameters can be dynamically modified, without stopping the cluster, even during runs.

At DPCC, several tests have been performed on a daily chain, to find a tradeoff between duration of the run and allocated resources: at some point, adding resources to a queue does not significantly shorten the processing (Figure 2), and it is more efficient to give these resources to other chains.

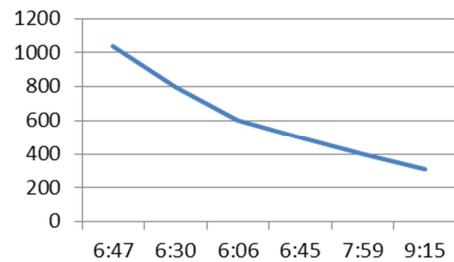


Figure 2: number of cores versus total processing duration of a daily chain.

4.2.2. Memory management

Initially, the configuration was carefully set up to share the cores. But after some months, it became clear that the main limiting resource was not the number of cores, but more the amount of memory.

To solve this issue, the number of reduces has been limited to ensure that each reduce has enough memory. The type of hardware has also been selected accordingly: the new computers ordered for cluster upgrades have a better storage capacity and more memory.

4.2.3. Cutting the datasets in chunks

Sometimes, for several reasons (timers, failures...), Hadoop tasks fail: after 3 failures the job itself is considered as failed, and all the computing hours are lost (the job must be resumed from the beginning, and some can take several weeks).

To reduce the consequences of such failures, the input data are cut into chunks: if the processing of a chunk fails, only this chunk has to be restarted. The chunk size is a tradeoff between the number of chunks to handle and the duration needed to process one chunk. The values are chosen at DPCC so that the computation of a chunk takes less than 24h.

Of course, this constraint has strong impacts on the design of the chains, and the chunk size is not the same for all the chains. Some chains cannot be cut in chunks.

4.2.4. Map-reduce deadlocks

Reduces use the outputs of maps. These outputs are stored locally on the nodes which ran the maps. When all the maps are finished, Hadoop launches the reduces. If a node fails at this time, some maps outputs become unavailable for the reduces. These maps should be run again to recompute the missing data. But meanwhile the reduces took all the resources, causing a deadlock.

The mechanisms provided by Hadoop to avoid this are not always efficient, and DPCC faced this situation several times. To avoid this, we define the maximum number of reduce executed at the same time to a value less than the number of cores allocated to the jobs. So, if a map shall be launch again, a set of core is always available to run it.

4.3. Lessons learned about hardware architecture

4.3.1. Taking care of the master nodes

The Hadoop cluster is managed by servers called master nodes. These nodes are crucial for the good health of the cluster, and in particular they are controlling HDFS.

In the initial architecture of the DPCC, these nodes were also hosting other services (Cassandra, ElasticSearch, centralized DPCC applications...). Sometimes, these applications overloaded the servers, causing failure of master nodes services.

During a major hardware upgrade of the cluster in June 2017, the other services were moved to new dedicated servers, leaving dedicated servers only with the master nodes. Since this upgrade, the platform has been far more stable.

4.3.2. High hardware availability

The Hadoop infrastructure is really robust to hardware failures. It is also possible to quickly commission or decommission some nodes in case of maintenance, without any impact on the processing.

4.3.3. Hardware scalability

The extension of the cluster can be done without interruptions of service. Hadoop can also manage heterogeneous machines from different hardware providers, taking advantage of technology improvements.

The cluster of DPCC takes advantage of this, with progressive yearly upgrades, adapted to a per-year budget and to increasing computation needs versus time.

But some side effects were noticed during the upgrades. When new nodes are added to the cluster, with nearly empty hard disks, Hadoop transfers data from old nodes to the new ones in order to balance the data storage over the cluster. In order not to saturate the network and not to take too much resources from the normal processing, this rebalancing can take a long time (several weeks). Moreover during this transition period the old nodes are highly loaded (because they host more data to process) and the new ones have nothing to do.

The lesson learned is even if commissioning new nodes is quick and easy, their full availability in the cluster is not immediate.

4.3.4. A validation cluster

As described above, the Hadoop cluster has many parameters (blocksize, queues, replication, chunk sizes...) and the choice of these parameters can have significant impacts on the performances.

So a validation cluster was set up at DPCC, with a smaller number of nodes but the same architecture. This validation cluster is used to test all the modifications before applying them to the production cluster.

It is also used to validate the scientific chains before running them on the real Gaia data.

The main drawback of this architecture is the availability of the data: because the real (and big) data are on the production cluster, subsets of filtered data need to be extracted and transferred to the validation cluster. This creates unwanted duplication of data and requires processing time on the production cluster.

For this reason, final chain tests (requiring large sets of real data) are performed on the production cluster.

4.3.5. Performances monitoring

Assessing the performances of the chains is a major concern for DPCC, in order to respect the calendar constraints described in §2, in particular the end of cycle deadlines.

But measuring the performances in such a distributed architecture is difficult. Hadoop gives a lot of statistics, but only at job level. End-to-end performances analyses are complex. Moreover, the processing duration is not a reliable measurement, as it strongly depends on the load of the cluster, with other chains running at the same time.

Dedicated tools have been developed in DPCC to help these analyses, but they still have to be improved.

5. PERSPECTIVES

Potential improvements of the DPCC architecture have been identified:

- upgrade to Hadoop V3,
- use Spark for some chains,
- use Dr. Elephant to monitor performances,
- re-organize the data in HDFS to improve access and filtering capacity.

6. CONCLUSION

After three years, the DPCC Hadoop architecture has proven its efficiency to be able to fulfill the Gaia big data processing challenges.

The pros are: management of huge data volumes, efficient parallel processing, efficient resource sharing, good reliability, scalability, robustness to hardware issues.

The cons are: a very complex configuration (requiring a dedicated team of experts), performances monitoring.

7. REFERENCES

- [DR1] Gaia Archive (Data Release 1), <https://gea.esac.esa.int/archive/>
 [DR Scenario] Gaia Data Release Scenario, <https://www.cosmos.esa.int/web/gaia/release>

BEYOND SENTINEL-2 WITH URTHE DAILY CONSTELLATION

Ramos, Jose Julio

Deimos Imaging, SLU, an UrtheCast Company

ABSTRACT

This paper introduces the UrtheDaily™ constellation as a Big Earth Observation Data complement of Sentinel-2 mission for scientific and commercial applications. It describes the constituent components from the spacecrafts and onboard optical payloads and communication systems, the carefully selected orbital geometry that -alongside the Ground Station Network distribution- optimizes the download data flow, the massive image production system running on the Amazon Web Services cloud, which serves products to UrtheCast's Kanvas platform and which will end up being part, as analytics-ready products, of third-party algorithms, systems, geoanalytics platforms and applications controlled and visualized from desktops, pads or cell phones.

As a conducting theme, and to give readers a reference, this paper compares UrtheDaily™'s offer with the existing Copernicus Sentinel-2 capabilities and study how users would benefit from a free and open data supply complemented with a compatible premium offer.

Index Terms— UrtheDaily™, Sentinel-2, Daily Revisit, Medium Resolution, Space Assets.

1. INTRODUCTION

UrtheCast is a leading provider of commercial satellite imagery and the ideal partner to meet current and future satellite imagery needs for geoanalytics. UrtheCast current space assets include two optical sensors mounted on the International Space Station (Theia and Iris) and two other optical sensors flying on Low Earth Orbits (Deimos-1 and Deimos-2). Constellations in development include UrtheDaily™ and OptiSAR™.

The future UrtheDaily™ constellation, formed by eight satellites, will provide scientific-quality multispectral images of the entire globe landmass minus Antarctica -every day- enabling actionable insights for the agriculture, forestry, and other industries. By leveraging Kanvas, our currently operational cloud-based distribution and exploitation system -powered by Esri's ArcGIS platform- UrtheDaily™ will make its data available in no more than 12 hours after acquisition.

Being designed with direct customer feedback and answering real business needs, it complements the Sentinel-2 open data services with compatible spectral bands but adding around 30 times more pixel density, on a daily basis, while guaranteeing businesses continuity with a Service Level

Agreement during the expected 10 years of operations and allowing access to data from areas as small as 5x5m, opening the way to applications never imagined before.

UrtheDaily™ is designed to acquire high-quality multispectral imagery of the entire Earth's landmass, at 5-m Ground Sample Distance (GSD), every day at 10:30 AM local time, always Nadir pointing. The system will be optimized for agricultural and change detection applications, enabling powerful geoanalytics capabilities that will provide enhanced value to the agriculture industry.

2. SPACE SEGMENT

The UrtheDaily™ space segment is a constellation of eight cross-calibrated, long-lifespan satellites, which has been optimized for daily coverage at all design trade-off decisions:

- Swath width vs. No. of satellites
- Swath width and GSD vs. Altitude
- Pixels per camera vs. No. of cameras
- Payloads vs. Power system design
- Bus size and weight vs. Launch system
- Spectral bands vs. Customers' needs
- Spectral bands vs. Other missions
- Ground Station Network (GSN) Antennas vs. Latency
- Daily throughput vs. Ground Segment architecture
- Automation vs. Operational costs

The spacecrafts, built by Surrey Satellite Technology Ltd. (SSTL) are based on the proven SSTL-250 bus, will have a mass of 340kg. Their dimensions are 1.1m x 1.1m x 0.8m and will have an expected lifespan of 10 years.

Payload sensors will always be imaging over land, and always be nadir pointing over a 360km swath. This configuration eliminates the need for on-demand payload tasking and simplifies operations.

Estimated launch date is on Q4 2019, when a SpaceX Falcon 9 will be used to put all spacecrafts at once on a Sun Synchronous Orbit at 600km of altitude, with 10:30 as local time at node. This carefully selected orbit guarantees global daily coverage: the world's landmass (minus Antarctica) acquired daily, with 140 million sq-km of multispectral imagery collected every 24 hours.

3. IMAGE PRODUCTION CHAIN

To achieve its mission objectives, the UrtheDaily™ production chain's design will apply many lessons learnt

from the Big Earth Observation Data engineering community as it will be:

- Deployed on the cloud leveraging Amazon Web Services (AWS) storage and computing capabilities/ AWS is the distinguished leader of the Infrastructure-as-a-Service sector [1], an established technological partner which will guarantee business continuity over the mission lifespan;
- Designed for scalability to comply with performance requirements (3 to 6 hours of imagery production latency) under variable data feed rates;
- Managed for optimizing operational costs choosing the best configuration of all chargeable resources, while always performing within constraints;
- Easily evolved, updatable and replicated for enabling system changes like, for instance, re-processing campaigns after algorithm tweaks and calibration adjustments;
- Fully automated reducing manual operations during collection, downlink, backhaul and data processing, cataloguing and delivery to users through UrtheCast's Canvas platform.

The UrtheDaily™ input data stream is more than 25 Terabytes of compressed raw data per day. The image production chain will decompress, correct, calibrate, annotate and convert this raw data in L1B data (pre-ortho), generating an estimated 86 Terabytes of data per day, a total of 314 Petabytes over 10 years of operations.

4. KANVAS DISTRIBUTION AND EXPLOITATION PLATFORM

The combined characteristics of the space and ground segment will make available L1B products within 12 hours from acquisition. Customers will access this data via Canvas, UrtheCast's platform powered by Esri's ArcGIS cloud technology. Esri is a supplier of geographic information system (GIS) software which in 2014 had approximately a 43 percent share of the GIS software market worldwide [2], more than any other vendor, and that in 2016 has grown to own "more than half of the market for GIS software, and its technology is used around the world by some 350,000 businesses, government agencies and NGOs" [3]. This company has been chosen for ensuring business continuity and to leverage their existing imagery platform capabilities. Therefore, Canvas will allow:

- On-the-fly ortho-rectification and indexes generation – human operators will obtain analytics-ready, L1C images on their screen within milliseconds;
- Geo-analytics - an ever-growing set of Earth Observation (EO) services and applications specifically designed for quick combination, analysis and extraction of information (including "Bring Your Own Algorithm", BYOA);
- Machine-to-machine interfaces including open Application Programming Interfaces (APIs, Open

Geospatial Consortium, OGC-compliant), data and metadata standards and format and Javascript and Python bindings;

- Visualisation - complete portfolio of visually captivating presentation options (including graphs, 2D- and 3D-maps of both raster and vector layers);
- GIS - close integration with unmatched Geographical Information Systems (GIS);
- Monetisation - the possibility to monetise developed services and products;
- Community - a massively big community of remote sensing experts, engineers and data scientists developing algorithms and final users consuming Earth Observation imagery and analytics;
- Vertical integration - seamless compatibility with technologies used for companies in much larger industries and sectors like forestry, agriculture, infrastructures, urban planning, defence and intelligence;
- Horizontal integration – connect third party applications, services and other platforms for leveraging legacy resources.

Canvas can be used in conjunction with existing Sentinel-2 data and services for operational needs, leveraging said machine-to-machine interfaces for integrating third-party applications and federating with existing and future platforms, like CKAN-based platforms, the Copernicus Data and Information Access Services (CDIAS, [4]) or geospatial data cubes, including national implementations of the Open Data Cube platform like the Australian Geoscience Data Cube (AGDC) [5].

Third-party platforms and application developers might choose five increasingly-efficient levels of integration with Canvas:

1. Local processing – Replicating data on alternatives infrastructure, moving the data to the user, i.e., obtaining complete product files and moving them to private infrastructures where they could be used by applications and services. Canvas supports a wide variety of open standards for raster images (including GeoTiff, JPEG, NetCDF and HDF), vector files (including shapefiles, GML, KML and GeoJSON), and metadata formats (including INSPIRE) [6].
2. Open APIs access – Accessing Canvas data and services like catalogue and on-demand processing and distribution capabilities through OGC services (including WMS, WFS, WMTS, WPS, WCS and CSW), REST GeoServices, OpenSearch end-points and others [6].
3. Apps development – Integrating provided programming libraries in your application code or quickly building WebApps configuring available builders. Developers may leverage JavaScript APIs for web development and native Software Development Kits (SDKs) for most development frameworks (including Android, iOS, Java, MacOS, .NET and QT) [7].

4. Platform extension – Automating geoprocessing workflows and extending provided capabilities with custom logic that can be executed in ArcGIS clients while leveraging existing scalability and security functions. Python bindings and ArcGIS Server Object Extensions (SOEs) and Server Object Interceptors (SOIs) are available to developers [8].

5. FOCUS ON ANALYTICS-READY DATA

UrtheCast has worked with geanalytics companies all over the world to refine the data specification to look to ensure that the UrtheDaily™ data will be information-rich, machine learning ready for multi-use applications straight from the Kanvas platform.

The spectral bands of the UrtheDaily™ constellation have been specifically selected to match Landsat-8, Sentinel-2, RapidEye and Deimos-1 bands to ease the constant/automatic in-flight cross-calibration with trusted references, minimize effects due to atmospheric variations, and to provide improved accuracy of key information products (e.g. Normalized Difference Vegetation Index, NDVI). The system is designed to provide a high Signal-to-Noise Ratio (SNR) and bit depth that goes a long way to reducing measurement uncertainty.

UrtheDaily™ will generate around 30 times more pixels per day than Sentinel-2 with similar data quality, which will be provided to customers as part of its “Imagery-as-a-Service” (IMaaS) function.

Regarding product types, while Sentinel-2 focuses on systematic generation of Level-1C (Top-Of-Atmosphere reflectances -TOA- in cartographic geometry) and proposes a prototype Level-2A (Bottom-Of-Atmosphere reflectance in cartographic geometry) to be generated on the user side [9], UrtheDaily™’s offer provides more options to customers:

- Level-1B products (radiometrically corrected imagery in TOA radiance values and in sensor geometry), including -but not applied- the refined geometric model which might be used to generate a L1C product. These products will be systematically generated and made available for distribution and exploitation.
- Level-1C products (plus indexes like NDVI) will be generated on-the-fly after demanded by human users of the Kanvas platform, using the provided-by-default geometric model, image reference databases and Digital Elevation Models (DEM) for ortho-rectification and grid projection, or customers may choose to use their own models, bases and grid systems which better fit their purposes. Level-1C product datasets will also be systematically generated to support higher-processing level products.

This IaaS function will potentially provide an additional catalogue of Value-Added Products, in order to integrate seamlessly into the wide Sentinel-2 user community. Products are:

- Level-2A: ρ BOA (Atmospherically-corrected product including cloud screening, and adjacency/slope effects correction).
- Level-2B: Generic Land Cover (with compatible classes matching available classification systems), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), Leaf Area Index (LEA), Fraction Vegetation Cover (FVC), Leaf Chlorophyll Content (C_{ab}) and Leaf Water Content (C_w).
- Level-3: Spatio-Temporal syntheses of Level 1C or 2A products.

While customers will have the liberty to choose their own auxiliary data for their final processing stages, by-default UrtheDaily™ products will use identical or compatible auxiliary and reference plus geometrical standards to ensure compatibility with Sentinel-2 products.

Cross-calibrated imagery with similar data quality, compatible product definitions and data being available as IMaaS in the cloud enables scientists and commercial Earth Observation developers to combine both free and public Sentinel-2 datasets with UrtheDaily™ products for multi-temporal, multi-resolution, multi-variable analysis.

6. CONCLUSIONS

This article has presented the planned UrtheDaily™ constellation by UrtheCast, as a Big Earth Observation Data system, designed as a complement to existing Copernicus Sentinel-2 mission, data and services.

UrtheDaily™ orbit and spacecrafts are designed following Sentinel-2’s approach of being capable of acquiring the whole world’s land masses (minus Antarctica) in a predictable orbit and always nadir-pointing. Existing capabilities are extended by providing daily revisits and obtaining 30 times more pixel density on each acquisition.

UrtheDaily™ ground segment will be prepared to acquire, process and store hundreds of Terabytes per day while serving data to customers in a few hours, all running over public clouds, allowing data and information sharing across the world. This system would generate hundreds of Petabytes of information during the expected 10 years life of the mission, ready to be used for long-time series analysis in the same fashion as the Copernicus missions.

UrtheDaily™ products are designed to be compatible in definition, present similar geometric and radiometric quality metrics and will be cross-calibrated with Sentinel-2’s “truth”. This ensures that users of both product datasets could seamlessly complement, combine and even substitute them in cases of no data availability.

UrtheDaily™ will distribute its products via the Kanvas platform, which offers multiple data exploitation capabilities and allows to integrate with third-party software applications and platforms, including Copernicus services.

7. REFERENCES

- [1] L. Leong, R. Bala, C. Lowery and D. Smith, “Magic Quadrant for Cloud Infrastructure as a Service, Worldwide”, Gartner, 15 June 2017.
- [2] “Independent Report Highlights Esri as Leader in global GIS market”, Esri, <http://www.esri.com/esri-news/releases/15-1qtr/independent-report-highlights-esri-as-leader-in-global-gis-market>, 2 March 2015.
- [3] “The Godfather of Digital Maps”, Miguel Helft, Forbes, <https://www.forbes.com/sites/miguelhelft/2016/02/10/the-godfather-of-digital-maps>. 29 February 2016.
- [4] “The upcoming Copernicus Data and Information Access Services (DIAS)”, Copernicus, <http://copernicus.eu/news/upcoming-copernicus-data-and-information-access-services-dias>, 26 May 2017.
- [5] <https://www.opendatacube.org/>
- [6] <http://www.esri.com/software/open>
- [7] <https://developers.arcgis.com/building-apps>
- [8] <https://developers.arcgis.com/extending-the-platform/>
- [9] “Sentinel-2 MSI product types”, ESA, <https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types>, obtained on 15 October 2017.

A NEW PARADIGM FOR THE EXPLOITATION OF THE SEMANTIC CONTENT OF LARGE ARCHIVES OF SATELLITE REMOTE SENSING IMAGES

Lorenzo Bruzzone¹, Manuel Bertoluzza¹ and Francesca Bovolo²

¹Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy

²Center for Information and Communication Technology, Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

This paper presents a new paradigm for extracting information from large databases of remote sensing images. It aims at improving any task applied to image time series by exploiting properties related to their temporal cross-dependence. Images part of the same time series are casually related to each other. As a consequence, the results of the tasks are mutually entangled. The proposed paradigm exploits this property and validates the results of the tasks one to each other to improve the overall performance. The paradigm is general and has relevant implications in Big Data analysis because it is suitable to archives containing not only Earth Observed images but any time-varying quantity or feature. Preliminary results show that change detection accuracy improves after the evaluation of the conservative property within the image time series.

Index Terms— Remote sensing, Big Data, data archives, change detection, data mining.

1. INTRODUCTION

The large number of new Earth Observation satellites available with improved revisit time and image resolution results in new challenges for the remote sensing community in terms of extraction of information from the data. For example, the Sentinel archive in 2016 experienced a daily publication rate of 4.58 TB/day and an annual growth of 250% [1]. This leads to a great potential information content that however is difficult to extract and exploit in a satisfactory and systematic way. There are several well-known challenges related to the processing of large archives, e.g., the capability to extract the rich semantic content from the data and the need to have computational architectures being able to efficiently process the data. In this paper, we discuss a completely different and new way to exploit efficiently entire archives of data. In greater detail, we present a novel paradigm for the information extraction from large databases that can be also useful for processing local images and validate the results against a larger archive. This paradigm, which is completely different from any previous idea exploited in Big Data (not only in remote sensing), addresses the information extraction by reformulat-

ing standard data analysis tasks, e.g., change detection (CD) or land-cover classification, in the framework of archive information exploitation. The main advantage of this framework is its intrinsic capability to improve the quality of standard products by exploiting the relationships among data and their properties in the archive. Moreover, it can be used for transferring knowledge from one image to any other one present in the database.

2. PROPOSED PARADIGM

Remote sensing images stored in large archives of Earth Observation missions can be interpreted as a huge set of long time series of the values assumed over time by each pixel at different spatial locations. The proposed paradigm exploits a set of intrinsic properties of these time series to improve typical processing schemes related to the analysis of single, pairs or sets of images, like classification, CD or trend analysis. The paradigm exploits the intrinsic cross-dependency of data (and the related products) in a image time series to constrain the results obtained on subsets of images (one, pairs or ensembles) with respect to other images in the archive. This can be achieved by generating auxiliary products for validating and improving the results of any elementary process. The paradigm is based on the following main properties.

Conservative Property: This property is fundamental for the proposed paradigm and allows us to relate the results of elementary processes applied to an image (or either a pair or an ensemble of images) with results obtained by the same process on other images acquired in the same year or same season. The images from the same time series are causally related with each other, so the results of any task must be consistent with each other. This enables the mutual validation of the results obtained within the time series. For example, let us consider the problem of binary CD in any pair of images acquired on the same area within the archive. According to the conservative property, the CD maps within the time series must be consistent with each other. In particular, for any pixel, in absence of errors in the bi-temporal CD between images, a change within any temporal loop (i.e., closed circular path) in the archive must be followed later in that loop by the opposite change so that the initial pixel state is guaranteed at

the end of the loop. Thus, an anomaly occurred in any change map along a loop can be found whenever the conservative property of the binary changes is not verified. Accordingly, an anomaly occurs on a pixel when the number of detected changes along a temporal loop is odd. A CD error can therefore be identified along loops that fail to satisfy the conservative property. When an anomaly has been identified, proper correction mechanisms can be used to fix CD errors between pairs of images within the time series.

Map of Transition States: The conservative property can be applied to each pixel of the images in the archive to generate a map of the transitions of the pixel state at different acquisition times, i.e., from one image to any another image in the archive. This can be achieved by applying the conservative property to many temporal loops to distinguish between reliable (permitted) and unreliable (non-permitted) paths. Let us consider again the problem of binary CD. If we consider the binary change as transition variable, it is possible to reliably estimate the state of the pixel status in one image following the changes occurred to pixels at the same location in the scene extracted from other images in the archive. This mechanism can be easily extended to estimate transitions among different classes if a multivariate rather than a binary CD technique is applied to the image time series. Given a pixel, the map of the transition potentially models relations among the states of the pixel in the full archive.

Flow of information along permitted paths: Given an image from the archive and the related products (or auxiliary information as ground reference data) we can exploit the map of state transitions to propagate the pixel-based information through the images of the archive. For example, if a reliable (permitted) path from an image to another is found, we can propagate the classification labels according to the map of the transition states. In particular, the information can flow only along a permitted path and be processed according to the state transition maps along that particular permitted path. This means that, for instance, if a pixel in a given image belongs to a class, we can easily propagate its label to the other images in the archive for which there are permitted paths associated with transition states that point out no changes. On the contrary, if the transition states point out changes along permitted paths the information that the label is changed can be propagated.

Accordingly, a general processing paradigm based on these three properties can be defined to enable the flow of information between different elements of the archive. This allows both the transfer of knowledge for enriching the information available at a given date and the validation and correction of errors on products obtained at local level. The next section introduces the proposed novel paradigm applied to the problem of binary CD in image time series followed by some preliminary results on the validation and correction of binary CD results in a synthetic and real archive of Earth Observed images.

3. APPLICATION TO THE BINARY CD PROBLEM

The proposed paradigm is used to define an iterative algorithm that defines the binary change variable within the graph theory in order to automatically cross-validate results obtained by any standard binary CD technique in the literature.

Let us define an image time series as an unordered set $\mathcal{T} = \{I_n\}$ containing N images acquired on the same scene at different times $t_n (n = 1, \dots, N)$. Any standard bi-temporal binary CD method f_p is applied to a pair of images I_i and I_j . This results in a CD map $\Delta_{i,j} = f_p(I_i, I_j)$ containing for each pixel p a boolean value corresponding to either the change (ω_c) or no-change (ω_n) label. The binary change variable is defined pixel-wise as a graph $\mathcal{G}_p(\mathcal{V}, \mathcal{E}, W)$ that comprises a set \mathcal{V} of N vertices and a set \mathcal{E} of M edges where W is the adjacency matrix. Vertices represent samples extracted from different images at the same pixel location p ; edges store the bi-temporal CD result obtained on the pair of images at pixel location p . Within the graph \mathcal{G} , a cycle is a closed simple path P in which head and tail vertices coincide: $C = \langle e_{i,j}, e_{j,k} \dots e_{j,i} \rangle$. For any path P of \mathcal{G} , let $\mu(P) \triangleq \sum_e w(e) \forall e \in P$ be the number of changes occurred along the path. The temporal consistency of the pixel status implies that the binary change variable is a conservative quantity. In absence of CD errors, along any closed circular path a change must be followed later by the same change in the opposite direction in order to reach the initial pixel status. Therefore, when changes are consistent, any cycle C in the graph \mathcal{G} must contain an even number of changes. The cycle consistency criterion can then be defined as: $\mu(C) \pmod{2} = 0 \forall C \in \mathcal{G}$. As a consequence, the concatenation of changes occurred along any path in \mathcal{G} must depend only on the initial and final vertex (i.e., path independence due to the conservative property).

The algorithm forces the temporal consistency of the changes within the time series by maintaining for all pixel positions of the scene p a graph with only temporally consistent cycles. The inconsistencies are removed by means of an iterative mechanism that evaluates all the images of the time series until the graph is connected. First, the iterative

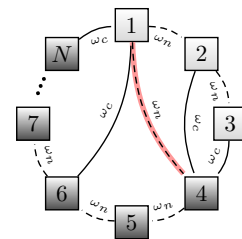


Fig. 1: Example of a consistent binary change variable as a graph where changes concatenated along different paths have the same result. No-change/change shown with dashed/full lines, respectively. Change $e_{1,4}$ is highlighted in red color is to be considered removed from the graph since identified by the proposed technique as an inconsistent change.

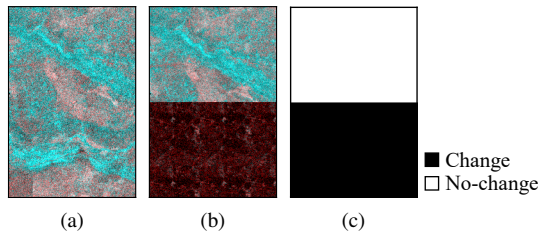


Fig. 2: (a) Pre-event and (b) post-event synthetic images with SNR = 18 dB. (c) Reference map (Synthetic dataset).

algorithm is initialized considering a connected subgraph \mathcal{G}_0 composed of a subset $V_0 \subset V$ of vertices such that any cycle in \mathcal{G}_0 is temporally consistent, i.e. the number of changes along them are even. Second, an increasing number of vertices are evaluated and all possible cycles in \mathcal{G} are considered until inconsistent changes are removed from the graph and \mathcal{G} is connected. Fig. 1 shows an example of graph representing a binary change variable containing changes detected within a time series of N images at the same pixel location. The iterative algorithm aims at removing all the inconsistent changes, e.g., edge $e_{1,4}$ that was highlighted in red color. After the removal of these inconsistencies of the conservative property, the graph contains only consistent changes. In the final result, i.e., after the removal of edge $e_{1,4}$, the concatenated changes along any path in the graph lead to the same pixel change status, i.e., either changed or no-changed. This allows us to test the presence of a change between a pair of vertices using any path connecting the two vertices. For example, the change occurred between vertices (1, 5) can be derived by integrating the changes along different paths: $\delta_{1,5} = \langle 1, 2, 4, 5 \rangle = \langle 1, 6, 5 \rangle = \langle 1, 2, 3, 4, 5 \rangle = \omega_c$.

The improved change detection map $\Delta_{i,j}^*$ is computed by considering the changes detected by the standard CD technique ($\Delta_{i,j}$) and correcting the change pixels that have been found by the proposed technique to be inconsistent with other changes within the time series. Due to the binary nature of the change variable, the correction of errors can be implemented by swapping the labels ω_n and ω_c . This technique captures step changes, i.e., the ones having more impact on the pixel intensity or color or spectral information, and builds temporally consistent change maps. Gradual changes along the time series are lost, but the higher temporal stability of the changes within the time series is preferred for consistency. For this reason, an appropriate temporal and spatial scale is required by the method for the improvement of the CD results in long image time series.

4. EXPERIMENTAL DATASETS

Experiments were conducted on two datasets composed of images acquired on the same scene at different times.

The first dataset is composed of 100 synthetic multispectral images of size 300×200 pixels with 2 spectral channels. They were generated starting from a pair of real multispectral images characterized by an abrupt change (Fig. 2). The change is artificially created between the first and successive

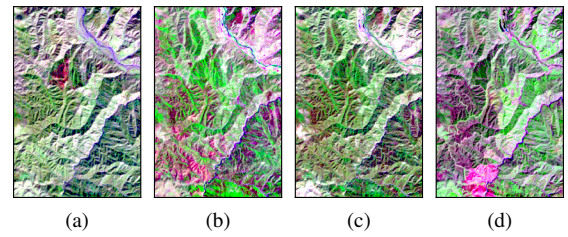


Fig. 3: False color representation (RGB=SWIR2,NIR,Blue) of four acquisitions extracted from the Landsat dataset composed of 812 images: (a) 31-Dec-82, (b) 28-Jul-98, (c) 13-Oct-06 and (d) 02-Apr-17.

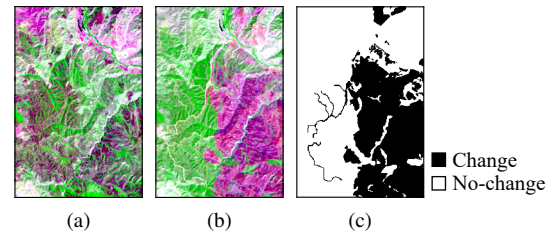


Fig. 4: False color representation (RGB=SWIR2,NIR,Blue) of the (a) pre- and (b) post-event images, acquired on 27-Jul-2015 and 04-Sept-2015, respectively. (c) Reference map (Landsat dataset).

acquisitions. The latter are replicas of the first real image where a given portion of the scene was replaced by a chunk of the second real acquisition. Then, uncorrelated pseudo-random White Gaussian noise characterized by a tunable SNR was added to each simulated image. To have realistic simulations, each image has a different noise realization [2].

The second dataset is composed of 812 co-registered multispectral images acquired by the Landsat-4, 5, 7 and 8 at a spatial resolution of 30m over an area located in Lake County, California. The images of size 300×200 pixels were co-registered and radiometrically corrected in surface reflectance and published in the Landsat Collections Tier 1 archive. Fig. 3 shows four of them in false colors. They were acquired between 1982 and 2017 (see Fig. 5). The experiment evaluates the CD results on a pair of images acquired before and after the Rocky and Jerusalem wildfires occurred during summer 2015 that destroyed approximately 7 122 hectares of forest. A reference map for the burned area is available for validation only. Fig. 4 shows the pair of images and its reference change map.

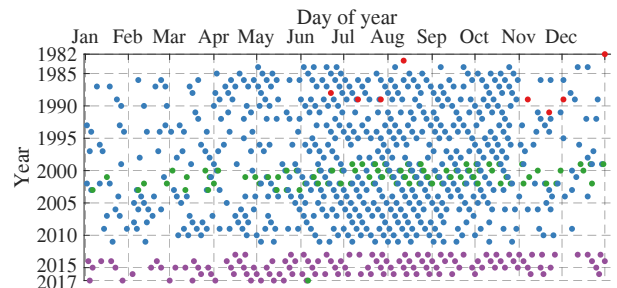


Fig. 5: Acquisition dates of the 812 multispectral images in the dataset acquired in California by Landsat mission (Landsat 4, 5, 7 and 8 as red, blue, green and purple markers, respectively).

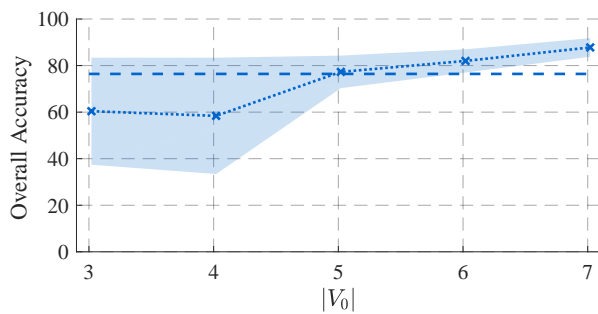


Fig. 6: Global overall accuracy obtained in the synthetic time series at different size of initial set of images ($|V_0|$). Bounded line shows the accuracy variability due to random initializations. Dashed line shows the global overall accuracy obtained by the standard approach.

5. EXPERIMENTAL RESULTS

A standard unsupervised bi-temporal CD method based on the Change Vector Analysis (CVA) has been applied to pairs of images of the two test time series. Change maps are obtained by thresholding the magnitude of the difference image according to well-established unsupervised and automatic strategies based on the minimum error Bayesian criterion [3]. In general, CD errors may be present in the bi-temporal CD results due to sensor noise or a sub-optimal estimation of the change/no-change class distributions performed by the automatic technique. We evaluate the improvements of the CD results using the proposed technique in the two datasets at a global level or considering a target pair of images, respectively.

In the first experiment, we considered the synthetic dataset that was specifically designed to be characterized by an abrupt change between the first and successive images. As a consequence, the reference changes occurred between any pair of images can be easily characterized. The assessment of the results obtained by the proposed technique is systematically done for all the pairs of images and all the pixels of the scene. It is based on the comparison of the change matrix obtained after applying the iterative algorithm and the reference change matrix. With this information, it is possible to compute the overall accuracy for the global image time series, considering all the pairs of images of the time series. In this experiment, the SNR of the noise added to the entire scene of 50 synthetic images is equal to 18 dB while the other half of images is characterized by a higher noise level, i.e., SNR = 12 dB. This setup aims at understanding the capability of the proposed approach to identify the location of CD errors within the time series. It is expected that they concentrate between synthetic images with a lower SNR.

The results of the first experiment are summarized in Fig. 6. The time-series CD accuracy is computed over all the pixels of the scene and all pairs of images within the time-series. The behavior of the global overall accuracy is shown at different values of the number of images ($|V_0|$) used to initialize of the iterative algorithm described in Section 3. With a larger initial set of images, the accuracy of the pro-

Table 1: Overall accuracies in the standard and proposed CD approaches using three different values for the target’s CD threshold.

	Target’s CD Threshold	Standard		Proposed	
		OA%	k	OA%	k
Bayesian	0.060	91.6%	0.82	92.4%	0.83
Underestimated	0.050	89.0%	0.77	93.1%	0.84
Overestimated	0.070	92.0%	0.83	92.5%	0.83

posed approach is more stable and better CD performance than the standard CD approach are reached. As expected, the initialization based on more images leads to a consensus of the different random executions of the iterative algorithm.

The second experiment assesses the effectiveness of the proposed technique for the correction of local CD errors in the real target pair characterized by a wildfire (Fig. 4). The analysis is performed by applying to the target pair the Bayesian decision threshold value for minimum error and simulated under-/over-estimated threshold values. The conservative property was evaluated pixel-wise along temporal loops within the time series to identify and correct possible CD errors in the target pair of images. A set of loops containing the target pair was used to determine how much reliable the changes in the target pair are. Accordingly, the most unreliable pixels in the target CD map have been selected and corrected by inverting their labels. Table I shows the comparison of the CD performance achieved on the target pair by the standard bi-temporal technique and the improved CD results based on the proposed iterative algorithm. The result points out the stability of the final accuracy achieved by exploiting the conservative property derived by other images in the archive that is the same irrespective of the change detection results obtained on the target pair. In other words, CD errors due to a poor decision threshold on the target pairs are recovered regardless of the initial CD performance of the bi-temporal technique.

6. CONCLUSIONS

In this paper, we presented a novel paradigm for the exploitation of the semantic content of large archives of satellite remote sensing images. These preliminary results show the validity of the proposed paradigm. Future work will apply the conservative property and related properties to Sentinel-2 image archive to: i) better characterize the potentials in improving CD performance, and ii) study the other properties related to the exploitation of the paradigm in the context of classification.

7. REFERENCES

- [1] ESA/Serco, “Sentinels Data Access Annual Report” *ESA report*, 2016.
- [2] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, “Sequential Spectral Change Vector Analysis for Iteratively Discovering and Detecting Multiple Changes in Hyperspectral Images” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4363–4378, 2015.
- [3] M. Zanetti, F. Bovolo, and L. Bruzzone, “Rayleigh-Rice Mixture Parameter Estimation via EM Algorithm for Change Detection in Multispectral Images” *IEEE Trans. Image Process.*, vol. 24, no. 12, 2015.

DETECTING ABNORMAL EVENTS IN MULTIVARIATE TELEMETRIES THANKS TO COVARIANCE ANALYSIS

C. Barreyre, B. Cabon, L. Boussouf

Airbus Defence and Space

B. Laurent, J-M. Loubes

Institut des Mathématiques de Toulouse

ABSTRACT

The space telemetries are relevant indicators to attest the health of the satellite. Any unexpected event that occur in one or more telemetries may be due to a misbehaviour of the satellite that must be detected. If a real anomaly appears, it is likely to affect more than one telemetry, changing the way the telemetries are structured in relation to each other. For this reason, we have developed a novel procedure to monitor groups of telemetries, based on a statistical test applied on the daily-covariance matrices. This multivariate approach can handle dimension reduction problems thanks to projections onto orthonormal functional bases.

Thanks to this method, we are able to early detect automatically divergences of some telemetries, and provide deep investigations when a real anomaly occurs.

Index Terms— satellites monitoring, outlier detection, multivariate analysis, covariance matrices, statistical tests, functional data analysis

1. INTRODUCTION

In order to monitor a satellite, thousands of telemetries are sent back to Earth almost real-timed. Most of these parameters are observed every thirty seconds, leading to consider terabytes of historical data. Some of them are followed up automatically thanks to experts knowledge, based on threshold rules under specific conditions. It is known that sometimes, abnormal events that occur in one or more telemetries may be due to a misbehaviour of the satellite that must be detected. This kind of work has already been treated by the ESA [1], the CNES [2] and the JAXA [3]. In their framework, the anomaly detection is done only on one single telemetry on which well-chosen features are computed to highlight anomalies.

However, if a real anomaly appears, it is likely to affect more than one telemetry, changing the way the telemetries are structured the ones with the others. Hence major events can sometimes generate subtle changes in several telemetries. That is why the analysis of the covariance between the telemetries is suitable for this purpose.

The ESA has already dealt with the multivariate aspect with DrMUST [4] software, which was designed for investigation

when it is known that a real anomaly occurred.

Our approach is to develop an unsupervised anomaly detection algorithm on multivariate telemetries based on covariance analysis. The covariance computation handles dimension reduction issues thanks to a functional approach.

This paper will be structured as follows. In the first section, we introduce the mathematical model, and explain how to compute the covariance matrix, in particular in a dimension reduction framework. In the second section, we introduce the statistical test to compare covariance matrices. In a last section, we apply this test on a real set of 20 telemetries.

2. MATHEMATICAL REPRESENTATION

2.1. Observation model

Suppose we have m daily-periodical telemetries X_j , where $j = 1, \dots, m$. Each telemetry is observed on p instants every day on n days. It means that, for each day i , we observe

$$X_{j,k}^{(i)} = f_j^{(i)}(t_k) + \varepsilon_{j,k}^{(i)}, \quad k = 1, \dots, p, \quad (1)$$

where $f_j^{(i)}$ is a continuous function in $\mathbb{L}^2([0, 1])$, and $\varepsilon_{j,k}^{(i)}$ is a Gaussian noise which variance is unknown. Without loss of generality, we assume that $t_k \in [0, 1]$, which is a time indicator within the day. If the telemetries are regularly sampled, then we can assume that $t_k = k/p$ days, for $k = 1, \dots, p$.

2.2. Covariance matrix

2.2.1. Empirical computation

In order to ease the notations, we introduce the computation of the covariance using a single day, leading to have $n = 1$. The covariance $\Gamma_{j,j'}$ between two telemetries X_j and $X_{j'}$ can be computed as

$$\Gamma_{j,j'} = \frac{1}{p} \sum_{k=1}^p (X_{j,k} - \bar{X}_j)(X_{j',k} - \bar{X}_{j'}),$$

where

$$\bar{X}_j = \frac{1}{p} \sum_{k=1}^p X_{j,k}.$$

One can also denote Y_j as the centered signal $Y_j = X_j - \bar{X}_j$, for all $j = 1, \dots, m$. Then, if $\mathbf{Y} = [Y_1, \dots, Y_m] \in \mathbb{R}^{p \times m}$, the covariance matrix $\mathbf{\Gamma}$ can be computed by

$$\mathbf{\Gamma} = \frac{1}{p} \mathbf{Y}^T \mathbf{Y}. \quad (2)$$

In this framework, we have one covariance matrix per day, that we denote $\mathbf{\Gamma}^{(i)}$, for each day $i = 1, \dots, n$.

Detecting some changes in the behaviour of the telemetries can be done by testing, for each day $2 \leq i \leq n$, the equality of the covariance matrices corresponding to the day i , $\mathbf{\Gamma}^{(i)}$ and $i - 1$, $\mathbf{\Gamma}^{(i-1)}$. It is also possible to consider more than one day of past telemetries, to build a more robust test. In this paper, the approach we handle consists in applying an existing non-parametric test based on the principal components comparison.

2.2.2. Reduced dimension

As the telemetries are observations of functional data, it is possible to represent the functions by using projections onto orthonormal bases. The functional approach can be in fact really relevant as soon as the sampling of the telemetries is not regular, or when some portions of data is missing, which is observed.

Given $f_j \in \mathbb{L}^2([0, 1])$, if $(\phi_\lambda)_{\lambda \in \mathbb{N}^*}$ is an orthonormal basis in $\mathbb{L}^2([0, 1])$, then the observed functions f_j can be decomposed in this basis,

$$f_j(t) = \sum_{\lambda \in \mathbb{N}^*} \theta_{j,\lambda} \phi_\lambda(t).$$

From the observations obeying to Model (1), the coefficients $\theta_{j,\lambda}$ can be estimated by their empirical counterparts

$$\hat{\theta}_{j,\lambda} = \frac{1}{p} \sum_{k=1}^p X_{j,k} \phi_\lambda(t_k).$$

This representation enables us to reduce the dimension of the data by considering a reduced number of coefficients, leading to consider the levels $\lambda \in \{1, \dots, d\}$, where $d \leq p$ is the number of components to retain.

Dimension reduction using projections is really common when we deal with functional data, see for example [5], [6], where many orthonormal bases are tested.

Let $\mathbf{\Phi} \in \mathbb{R}^{p \times d}$ be the matrix representation of the functional basis, where $\mathbf{\Phi}_{\lambda,k} = \phi_\lambda(t_k)$, and $\mathbf{\Theta} \in \mathbb{R}^{m \times d}$ is the matrix of the estimated coefficients of \mathbf{Y} in this basis, where $\mathbf{\Theta}_{j,\lambda} = \hat{\theta}_{j,\lambda}$, then we can approach \mathbf{Y} by $\tilde{\mathbf{Y}} = \mathbf{\Phi} \mathbf{\Theta}^T$. The approached covariance matrix $\tilde{\mathbf{\Gamma}}$ can be computed by applying the equation (2) to $\tilde{\mathbf{Y}}$, and if $(\phi_\lambda)_{\lambda \in \mathbb{N}^*}$ is orthonormal in \mathbb{R}^p , then $\mathbf{\Phi}^T \mathbf{\Phi} = \mathbf{I}_d$, and we get

$$\tilde{\mathbf{\Gamma}} = \frac{1}{p} \mathbf{\Theta} \mathbf{\Theta}^T. \quad (3)$$

This property enables to use dimension reduction for computing the covariance.

2.2.3. Example

We consider a set of $m = 50$ simulated curves sampled on 256 points. We choose to reduce the dimension of the curves thanks to a Haar wavelet basis, as this basis is orthonormal both in $\mathbb{L}^2([0, 1])$ and \mathbb{R}^p . Let us first recall its definition. We set $\psi = \mathbb{1}_{[0,1/2[} - \mathbb{1}_{[1/2,1[}$. For all $l \geq 0$, $k \in \Lambda(l) = \{0, 1, \dots, 2^l - 1\}$, let $\phi_{l,k}(x) = 2^{l/2} \psi(2^l x - k)$. The functions $(\phi_0, \phi_{l,k}, l \geq 0, k \in \Lambda(l))$ form the orthonormal Haar basis of $\mathbb{L}^2([0, 1])$. The index l is recalled as the scale, and k as the position.

For this application, we retain the $(2^{L+1} - 1)$ levels corresponding to the wavelet levels for which $0 \leq l \leq L$, for a fixed upper scale L . If we represent the observations before and after dimension reduction, with $L = 4$ as upper scale, we obtain the results represented in Figure 1.

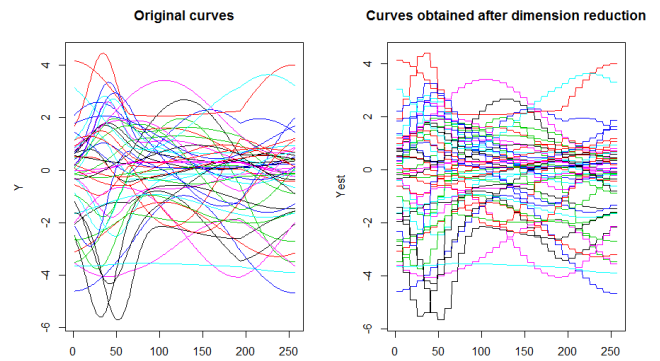


Fig. 1. Dimension reduction with Haar wavelets, where only the levels up to $L = 4$ are selected. Initial functions (on the left) are sampled on $p = 256$ time stamps, whereas the reduced dimension curves (on the right) are obtained from $d = 31$ coefficients.

In order to get the sensitivity of the dimension reduction on the error on the covariance matrix, we compute the covariance matrix from the raw-data $\mathbf{\Gamma}$ by using the equation (2), then we compute $\tilde{\mathbf{\Gamma}}_L$ by using the dimension reduction from the equation (3). In fact, we can estimate the relative error ϵ_L between the two covariance matrices by

$$\epsilon_L = \sqrt{\frac{1}{m^2} \sum_{j,j'=1}^m \left(\frac{\mathbf{\Gamma}_{j,j'} - (\tilde{\mathbf{\Gamma}}_L)_{j,j'}}{\mathbf{\Gamma}_{j,j'}} \right)^2}.$$

From the example presented in figure 1, we get $\epsilon_4 = 0.052$, which represents a relative error which is really small. With $L = 5$, then the error decreases to $\epsilon_5 = 0.013$, whereas $\epsilon_3 = 0.2$ for $L = 3$. For this application we should keep $L \geq 4$. The covariance computed after dimension reduction is really close to the covariance computed on the raw-data as soon as the dimension reduction is not too hard.

3. UNIVARIATE COVARIANCE TEST

Suppose we have only two days $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ of a daily-periodical telemetry that we want to compare, under the hypothesis that the instants $\mathbf{Y}_{:,k}^{(1)}$, $k = 1, \dots, p$ are i.i.d. The daily-periodical assumption is justified by the natural periodicity of the geostationary satellites. We would like test the following hypothesis :

$$H_0 : \{\mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}^{(2)}\}. \quad (4)$$

We choose to apply a non-parametric test introduced by Fremdt et al. [7] that was primary defined to test the equality of the covariance structures in two functional samples. The authors noticed that testing the equality of the covariance is equivalent to test that both samples have the same functional principal components. We adapted it to our problem, where the functional samples here are two days of periodical telemetries. The test consists into the following steps. Other tests inspired by the principal components comparison exist, as in [8].

- Let us consider a reference matrix $\mathbf{\Gamma}$. It can be for instance the mean matrix between the two considered covariance matrices.

$$\mathbf{\Gamma} = \frac{1}{2}(\mathbf{\Gamma}^{(1)} + \mathbf{\Gamma}^{(2)}).$$

One can also consider the covariance matrix computed on the barycenter distribution in the Wasserstein space, as it is proposed by Le Gouic et al. [9].

- Let φ_λ , $\lambda = 1, \dots, d$ be the d first ordered eigen vectors obtained by diagonalizing $\mathbf{\Gamma}$, and $a^{(1)}, a^{(2)} \in \mathbb{R}^{p \times d}$ be the coefficients obtained by projecting the telemetries from both days in this basis. They can be computed by

$$a^{(1)} = \mathbf{Y}^{(1)} \boldsymbol{\varphi} \text{ and } a^{(2)} = \mathbf{Y}^{(2)} \boldsymbol{\varphi}.$$

They are the d first principal components characterizing the sampling instants, and we assume that the eigen values $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)$ are strictly decreasing.

- The matrix $\Delta = \boldsymbol{\varphi}^T \mathbf{\Gamma} \boldsymbol{\varphi} = \boldsymbol{\lambda} I_d$ is the eigen values matrix. Under H_0 , we have $\Delta = \Delta^{(1)} = \Delta^{(2)}$, where $\Delta^{(1)}$ (resp. $\Delta^{(2)}$) can be computed by

$$\Delta^{(1)} = \boldsymbol{\varphi}^T \mathbf{\Gamma}^{(1)} \boldsymbol{\varphi} = \frac{1}{p} (a^{(1)})^T a^{(1)} \in \mathbb{R}^{d \times d}.$$

- Denote $\xi = \text{Vech}(\Delta^{(2)} - \Delta^{(1)})$, which is the vectorised version of the matrix $(\Delta^{(2)} - \Delta^{(1)})$ where we keep only the indexes corresponding to the upper triangle matrix. Consequently, we have $\xi \in \mathbb{R}^{\frac{d(d+1)}{2}}$. The test (4) is equivalent to test the hypothesis $H_0 : \{\xi = 0\}$.

- Denote $L \in \mathbb{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}}$ as the covariance matrix of ξ . The computation details are provided in [7].
- Then, it can be shown thanks to [7] that, under the null hypothesis H_0 the statistics

$$T = p \xi^T L^{-1} \xi \rightarrow \chi_{\frac{d(d+1)}{2}}^2 \text{ as } p \rightarrow \infty.$$

- The hypothesis is then rejected at a level α if

$$T \geq \chi_{d(d+1)/2, 1-\alpha}^2$$

where $\chi_{d(d+1)/2, 1-\alpha}^2$ is the $1 - \alpha$ level of a chi-squared distribution with $\frac{d(d+1)}{2}$ levels of freedom.

We can then apply this test every two-consecutive days in order to catch the daily changes in the covariance structure.

4. APPLICATION ON A REAL SET OF TELEMETRIES

We choose to apply the test on a group of 20 correlated telemetries that are closely linked. Those telemetries are daily periodical and can be cut into days. We have observed those telemetries on 365 days, then we can apply 364 tests, for each couple of days $(i-1, i)$, for $i = 2, \dots, n$, to catch the changes in the behaviour of those telemetries. We have reported in the following plot the test statistics T_i computed for each new day $i = 2, \dots, n$, represented in Figure 2. The days where the test is rejected at the level 5% are represented in red.

We can see that one statistic is much greater than the other

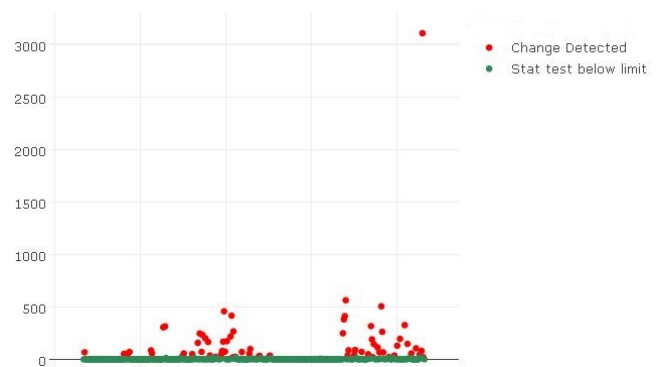


Fig. 2. Test statistics and events detected

values. We plot the telemetries, represented on Figure 3 to understand why the test was rejected so strongly. A red background has been added to identify the days corresponding to the detected change. With a human eye, we can see that the telemetries exhibit three main changes of behaviour at the same time.

These events changed strongly the covariance structure, and all the telemetries were impacted by this event. The upper

telemetries are the ones for which the telemetries start to decrease that day, not necessarily simultaneously. The telemetries in the center exhibit only light pattern changes. The telemetries in the bottom of the figure are the telemetries that increase. If we look at some other events, we are always able

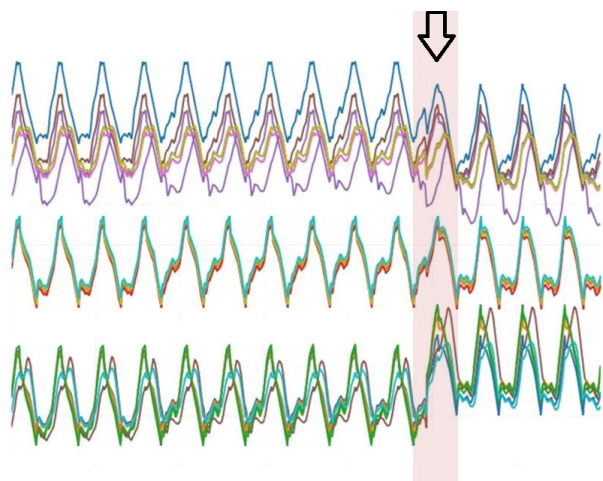


Fig. 3. Behaviour of the telemetries around the day matching to the highest statistical value

to note some changes by eye. For example, the figure 4, corresponding to the second highest value of the statistical test, shows that the black telemetry is less correlated to the other group of telemetries from the given day. This type of events is less evident to see by eye, and logically less evident to catch with hand-made processings. The higher the test statistics is, the more evident are the events. However, subtle changes can also be caught by this algorithm.

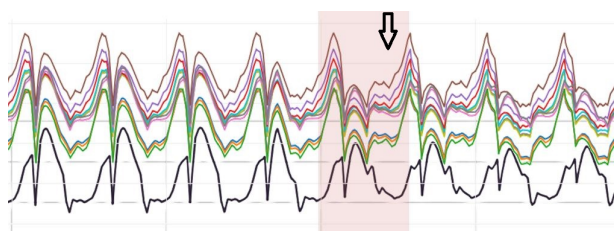


Fig. 4. Another anomaly detected : the black signal does not have the same shape as the other signals from the day represented with the red background

5. CONCLUSION

We have seen that we are able to detect events that change the behaviour of groups of telemetries. The method we propose is really fast, and it can be applied for a large number of telemetries. To follow-up the satellite in-flight, such methods can be applied on clusters of telemetries, for the daily monitoring as well as supporting deeper investigations. Coupled

with univariate methods, this algorithm is really efficient in highlighting the most abnormal behaviours almost real-timed. Those methods are already implemented within a web application at Airbus Defence and Space.

6. REFERENCES

- [1] José-Antonio Martínez-Heras, Alessandro Donati, Marcus GF Kirsch, and Frederic Schmidt, “New telemetry monitoring paradigm with novelty detection,” in *SpaceOps 2012 Conference, Stockholm, Sweden, 2012*, pp. 11–15.
- [2] Sylvain Fuertes, Gilles Picart, Jean-Yves Tourneret, Lotfi Chaari, André Ferrari, and Cédric Richard, “Improving spacecraft health monitoring with automatic anomaly detection techniques,” in *14th International Conference on Space Operations*, 2016, p. 2430.
- [3] Ryohei Fujimaki, Takehisa Yairi, and Kazuo Machida, “An approach to spacecraft anomaly detection problem using kernel feature space,” in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005, pp. 401–410.
- [4] José-Antonio Martínez-Heras, Alessandro Donati, Bruno Sousa, and Jörg Fischer, “Drumsta data mining approach for anomaly investigation,” in *12th International Conference on Space Operations*, 2012, pp. 11–15.
- [5] Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, and Jean-Michel Poggi, “Clustering functional data using wavelets,” *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 11, no. 01, pp. 1350003, 2013.
- [6] Benjamin Auder and Aurélie Fischer, “Projection-based curve clustering,” *J. Stat. Comput. Simul.*, vol. 82, no. 8, pp. 1145–1168, 2012.
- [7] Stefan Fremdt, Josef G Steinebach, Lajos Horváth, and Piotr Kokoszka, “Testing the equality of covariance operators in functional samples,” *Scandinavian Journal of Statistics*, vol. 40, no. 1, pp. 138–152, 2013.
- [8] Ioana Ilea, Lionel Bombrun, Christian Germain, Romulus Terebes, and Monica Borda, “Statistical hypothesis test for robust classification on the space of covariance matrices,” in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 271–275.
- [9] Thibaut Le Gouic and Jean-Michel Loubes, “Existence and consistency of Wasserstein barycenters,” *Probability Theory and Related Fields*, pp. 1–17, 2016.

CLOUD APPROACH TO AUTOMATED CROP CLASSIFICATION USING SENTINEL-1 IMAGERY

Andrii Shelestov^{1,2,3}, Mykola Lavreniuk^{1,2,3}, Andrii Kolotii^{1,2,3}, Vladimir Vasiliev³, Leonid Shumilo^{1,2,3},
Nataliia Kussul^{1,2}

1 Space Research Institute NASU-SSAU, Kyiv, Ukraine

2 National Technical University of Ukraine “Igor Sikorsky Kiev Polytechnic Institute”, Kyiv, Ukraine

3 EOSDA, Kyiv, Ukraine

ABSTRACT

For accurate crop classification, it is necessary to use time-series of high-resolution satellite data to better discriminate certain crop types. This task brings the following challenges: large amount of satellite data for download, Big data processing and computational resources for utilization of the state-of-the-art classification approaches. For solving these problems, we have developed an automated crop classification system CropZoom which is based on machine learning and deep learning techniques. By deployment of the system on the cloud platform, we can overcome challenges of Big data downloading and processing. In this paper, we present system architecture and describe the experiments on structural and parametric identification of machine learning models utilized in the system.

Index Terms— cloud platform, machine learning, crop classification, satellite images

1. INTRODUCTION

During last years a new era of free satellite data has started. With the launch of Sentinel-1 (both A and B) synthetic-aperture radar (SAR) data which are weather independent and freely available, new opportunities for crop classification have opened.

Crop mapping based on high resolution satellite data is a very important component for solving a number of applied problems, in particular crop area estimation [1] – [2], yield forecasting [3] – [4] and drought risk quantification [5]. Earlier, SAR data were quite expensive, infrequent and most of the studies on crop state assessment and crop type mapping were performed with optical data only. Due to clouds and shadows, the amount of available optical data over the region of interest is limited. High spatial (10 m) and temporal (6 days revisit) resolution of the Sentinel-1 mission bring new opportunities in the agriculture domain and challenges of “Big data” problems (for Ukraine more than 10 Tb per year) in Remote Sensing that should be addressed.

The main problems of dealing with big amount of SAR data for large-scale areas, such as Ukraine, are the following: time-consuming data downloading (bandwidth),

computationally intensive SAR data preprocessing (HPC), and storage capacity consuming (storage).

We have developed an automated crop classification system CropZoom which is based on machine learning and deep learning techniques. By deployment of the system on the cloud platform, we can overcome challenges of Big data downloading and processing. For instance, Amazon platform provides easy and fast access to Sentinel -1, -2 imagery via scalable Amazon Web Services (AWS) infrastructure (both storage S3 and computational instances EC2) to solve Big Data downloading and storing problem and also provides powerful computational resources which are necessary for running advanced deep learning and machine learning methods.

It is difficult to identify optimal architecture of the machine learning model and adjust its parameters locally to provide the best overall accuracy of classification due to large variety of parameter combinations [10] – [11]. Therefore, the main emphasize of this research is done on the methodology for structural and parameter identification of machine learning model for crop classification in a cloud environment and deploying classificatory into the cloud platform for large scale crop mapping.

2. SYSTEM ARCHITECTURE

To solve crop classification and crop mapping problem in near real time we need time series of satellite imagery and in-situ data for training, validation and test sets. Since clouds are the common problems for Ukraine and many other European countries, optical images are usually not enough for reliable classification. Therefore, for crop classification in these case studies, we utilized time series of Sentinel-1 SAR data. We collected in-situ data for study areas with ground surveys and digitized them as vector polygons or geo-referenced points with crop type labels.

To decrease the time for satellite imagery downloading it makes sense to deploy a classification system in the cloud environment, for example, Amazon, where Sentinel-1 data are already available for free. An architecture component diagram of the proposed classification system CropZoom is shown in Fig. 1.

The main operations in data processing chain are: satellite images downloading and preprocessing, in-situ data

collection, classification itself and validation, crop specific maps visualization and delivery to the end users.

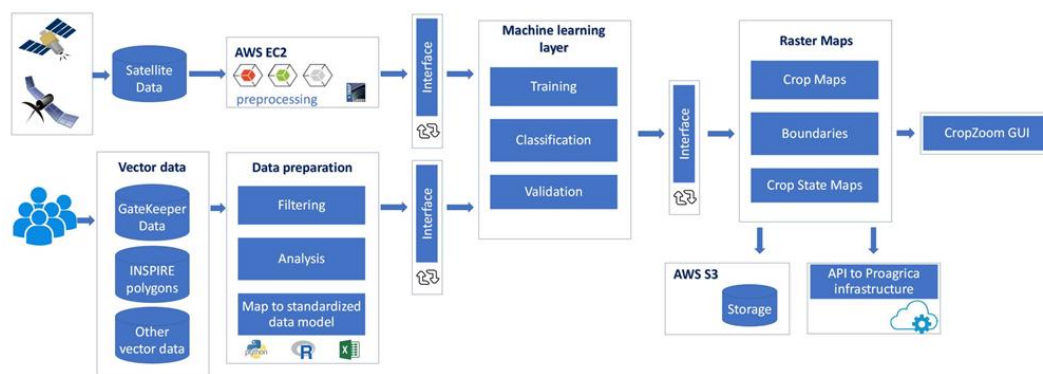


Fig. 1. Typical architecture component diagram for crop classification CropZoom system

3. CLASSIFICATION METHODOLOGY

Land cover and crop classification in the system is being done with three most popular and most accurate machine learning techniques: Random Forest (RF) [6], Support Vector Machine (SVM) [6], [7] and Artificial Neural Network (NN) [7] – [9]. Each method has its own advantages and disadvantages. Crop classification map accuracy often depends not only on machine learning method but also in-situ data quality and properly selected parameters of classifier.

It is not possible to create a universal crop classification model for any territory due to different nomenclature of crops to be grown at the territory and different number and of satellite data available. Therefore, dimension of input and output of the classifier could vary for different areas. Parameters for machine learning model could vary due to different number of input images as a result different number of input features, different number of in-situ samples, and percentage of the present noise in training data. So it is difficult to identify a generic classification model for all available data. For solving this problem, we propose to train the most accurate three classifiers: RF, SVM and NN and select the best one for crop mapping in different study areas independently. To provide the robustness of our methods we considered Ukraine and England as study areas. For this purpose, the best hyper-parameters grid search approach with a K-fold cross-validation scheme was used. The performance of the selected hyper-parameters and trained model is estimated on an independent evaluation set that was not used during the model selection step.

4. RESULTS

4.1. Structural and parameter identification of machine learning models

For this experiment, we utilized data from ESA Sentinel-1 SAR satellite. Time-series consist of approximately 45 images during the vegetation season for each location (more than 1000 images for Ukraine). Each SAR image has VV and VH polarizations, so for each pixel the length of feature vector equals 90. For training the classifier and final validation we have two different independent sample sets. Each set consists of 6428 samples.

For the SVM the most sensitive parameters for classification accuracy are: gamma, C and kernel type. In this experiment, we investigated two the most common types of kernel: radial basis function (RBF) and sigmoid (Fig.2, upper and bottom pictures respectively). As a result, RBF kernel is more appropriate for crop classification tasks. Moreover, it is not necessary to do exhausted grid search within hundred parameters possibilities, because the highest accuracies lies on the diagonal $(10^{-3}, 10^2)$, $(10^{-2}, 10^3)$, $(10^{-1}, 10^4)$, $(1, 10^5)$ for gamma and C, respectively. There is an explanation for such results. If gamma parameter is big (more 10^2), the impact area of the support vectors includes only the support vector itself and it leads to overfitting. Even fine-tuned C parameter will not be able to prevent overfitting. Otherwise, when gamma is very small, the model is too weak and cannot fit the complex data with high enough accuracy. In addition, it should be noted that there is some clandestine rule that recommends picking gamma coefficient nearby inverse value of the input feature vector size. Based on these results we can avoid grid search at all.

For the MLP method, we discovered sensitivity of classification accuracy from such parameters as alpha coefficient for regularization and number of hidden neurons. As it is shown in Fig. 3, the number of neurons has much stronger impact on the overall accuracy comparing to the

regularization coefficient. So, the best way for determining alpha is 10^{-2} or 10^{-3} for preventing an overfitting. Generally, how to found the most appropriate number of hidden neurons depends on each situation independently. However, there are some empirical rules: it should never be more than twice as large as the input layer or (number of inputs + outputs) * (2/3). Therefore, it is essential to use grid search approach for founding the best number of hidden neurons in this diapason.

The third classifier being investigated is Random Forest. The RF classifier has been discovered for it maximum depth for each decision tree and number of trees in the forest (Fig. 4). We observe that higher number of trees provides us better accuracy, at the same time makes our program 5 times slower. The recommended number of the trees in RF is 100, after this value the accuracy does not improve significantly. However, using the capability of powerful machine you could picking up it higher. The situation with maximum depth of each tree is similar but has its own specific. Deeper tree leads to decreasing an error nevertheless starting from some point it begins overfitting the training data. Depends on the noise presents in the training data we recommend to define maximum depth as 200 for data almost without noise or 100 in other case.

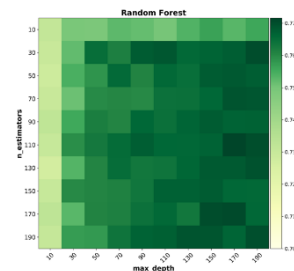


Fig. 4. Best hyper-parameters grid search for RF method.

4.2. Along the season crop classification maps

Utilizing the proposed CropZoom system, we have possibility to easily renew the obtained crop classification map with each new available Sentinel-1 image. For some tasks, it is essential to use crop mask or crop classification map before the end of the vegetation season, for instance, for yield forecasting. In Fig. 5, it is shown dependencies of the overall accuracy of crop classification maps during the vegetation period and number of utilized images. We can observe strong correlation between overall accuracy and number of image and in the end of the season we have the most accurate map.

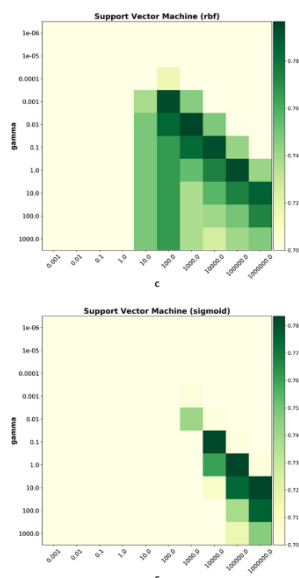


Fig. 2. Best hyper-parameters grid search for RBF (upper) and sigmoid (bottom) kernels in SVM classifier.

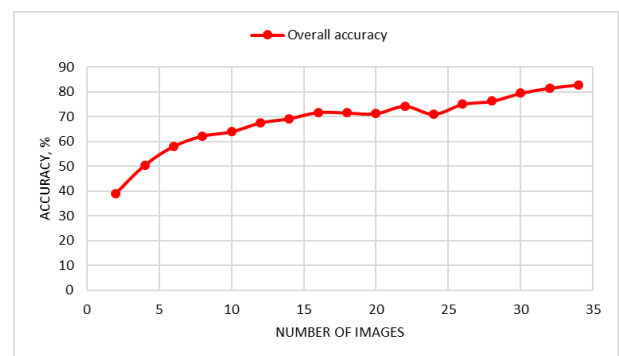


Fig. 5. Crop classification accuracy during the vegetation period depending on the number of scenes acquired (revisit time – 6 days).

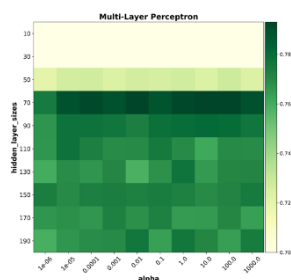


Fig. 3. Best hyper-parameters grid search for MLP method.

In Fig. 6 it is shown crop classification map for Ukraine territory for 2017 in the end of vegetation period. There are 13 classes and the overall accuracy was 89.7.

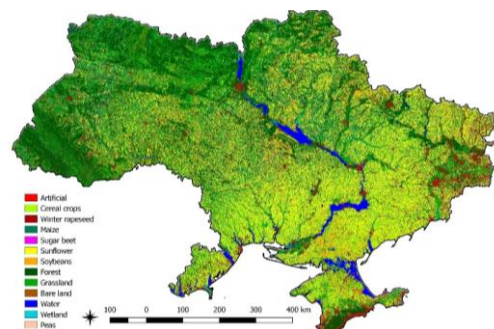


Fig. 6. Crop classification map for national use case – Ukraine for 2017.

For each class the F1-score has been calculated based on independent test set: artificial - 63.9, cereal crops - 95.7, winter rapeseed - 96.6, maize - 81.8, sugar beet - 78.5, peas - 93.7, sunflower - 94.4, soybeans - 57.6, forest - 96.9, water - 98.2, grassland - 59.2, bare land - 49.3, wetland - 70.7.

5. ANALYSIS OF SYSTEM EFFICIENCY

Let's consider the benefits of system implementation in cloud environment.

In case of such test site as Ukraine full coverage (9 paths of S1) for vegetation period of single year consists of more than 800 scenes (with S1A only). Downloading such amount of data from ESA SciHub on relatively small speed is time consuming (up to 2 full days on average speed of 5 mb/sec with storing within local infrastructure) while corresponding data download for preprocessing on Amazon from Alaska Space Facilities will take 10-15 hours (with average speed 15-20 mb/sec). Therefore, cloud-based implementation of system allows shortening time to access satellite data in 5-10 times.

Vector data collection and preparation is not computationally intensive operation. At the same time, it requires human resources for in-situ data preparation and open-source data collection (source of information for raster maps creation). So, a local solution for this operation is preferable.

SAR data preprocessing includes filtration (speckle reduction), calibration, orthorectification and terrain correction. It takes comparable time on local resources and on Amazon EC2 instance (per single path) but computational resources scaling in the cloud allows us to utilize many instances in parallel for preprocessing each path on separate instance simultaneously with corresponding speed-up of processing. So, cloud implementation is more preferable for satellite data preprocessing.

Cloud environment provides more opportunities to utilize advanced machine learning techniques, since they require more computational resources for efficient implementation [6]. In the case of local implementation, data download from the cloud is required for crop type classification and its duration is comparable with crop mapping time. That is why cloud based implementation is preferable for machine learning.

6. CONCLUSIONS

The main advantages of cloud-based approach for dealing with classification of big amounts of satellite data are: high speed of data downloading, parallel data utilization for covering big area on different instances simultaneously (data parallelism), powerful computational resources which are required for advanced deep learning and machine learning methods, large storage capacity. Utilization of the cloud platform allows significant decreasing of computational time for training classifier with best parameters and time for obtaining crop classification map.

This allowed us to train models on large amount of training data and optimize through large number of parameters compared to local machine.

Consequently, it allows us frequently renew (at least every 12 days) crop type maps during the vegetation season with higher accuracy based on time series of satellite imagery.

7. REFERENCES

- [1] J. Gallego, et al., "Efficiency assessment of different approaches to crop classification based on satellite and ground observations," *J. of Auto. and Inf. Sciences*, vol. 44, no. 5, 2012.
- [2] F. J. Gallego, et al., "Efficiency assessment of using satellite data for crop area estimation in Ukraine," *International Journal of Applied E. O. and Geoinformation*, vol. 29, pp. 22-30, 2014.
- [3] F. Kogan et al., "Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 23, pp. 192-203, 2013.
- [4] A. Kolotii, et al., "Comparison of biophysical and satellite predictors for wheat yield forecasting in Ukraine," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XL-7/W3, pp. 39-44, 2015. DOI: 10.5194/isprsarchives-XL-7-W3-39-2015.
- [5] S. Skakun, N. Kussul, A. Shelestov, and O. Kussul, "The use of satellite data for agriculture drought risk quantification in Ukraine," *Geomatics, Natural Hazards and Risk*, vol. 7, no. 3, pp. 901-917, 2016.
- [6] A. Shelestov, M. Lavreniuk, N. Kussul, A. Novikov, and S. Skakun, "Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping," *Front. Earth Sci.*, vol. 5, no. 17, pp. 1-10, 2017. doi: 10.3389/feart.2017.00017.
- [7] F. Waldner, et al., "Towards a set of agrosystem-specific cropland mapping methods to address the global cropland diversity," *International Journal of Remote Sensing*, vol. 37, no. 14, pp. 3196-3231, 2016.
- [8] N. Kussul, G. Lemoine, F. J. Gallego, S. V. Skakun, M. Lavreniuk, and A. Y. Shelestov, "Parcel-Based Crop Classification in Ukraine Using Landsat-8 Data and Sentinel-1A Data," *IEEE J. of Select. Topics in Appl. Earth Observ. and Rem. Sens.*, vol. 9, no. 6, pp. 2500-2508, 2016.
- [9] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778-782, 2017.
- [10] S. Fritz et al., "The need for improved maps of global cropland," *Eos, Transactions American Geophysical Union*, vol. 94, no. 3, 31-32, 2010.
- [11] Y. Ma, et al. "Remote sensing big data computing: challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47-60, 2015.

SPATIO-TEMPORAL ANALYSIS OF CHANGE WITH SENTINEL IMAGERY ON THE GOOGLE EARTH ENGINE

Morton J. Canty

Heinsberger Str. 18
D-52428 Jülich, Germany

Allan A. Nielsen

Technical University of Denmark
Applied Mathematics and Computer Science
DK-2800 Kgs. Lyngby, Denmark

1. INTRODUCTION

A characteristic task in remote sensing Earth observation involves the registration of changes which may signal environmentally significant events. The Sentinel-1 synthetic aperture radar (SAR) and the Sentinel-2 optical/visible-infrared space-borne platforms, with spatial resolutions of the order of 10-20 meters and revisit times of the order of days, provide an attractive source of data for change detection tasks, the SAR imagery especially providing complete independence from solar illumination and cloud cover. A convenient source of such data is the Google Earth Engine which gives near real time data access and which has an application programming interface for the access and for the processing the data. Here we make open-source automatic change detection software and for optical data also automatic radiometric normalization software available.

2. CHANGE DETECTION IN SAR DATA

In [1] a change detection procedure for multi-look polarimetric SAR data [2] is described involving a test statistic (and its factorization) for the equality of polarimetric covariance matrices following the complex Wishart distribution. The procedure is capable of determining, on a per-pixel basis, if and when a change at any prescribed significance level has occurred in a time series of SAR images. Single polarization (power data, dimensionality $p = 1$), dual polarization (for example vertically polarized transmission, vertical and horizontal reception, $p = 2$) and full or quad polarization (all four combinations of vertical and horizontal transmission/reception, $p = 3$) can be analyzed.

The term multi-look in SAR imagery refers to the number of independent observations (termed the equivalent number of looks, ENL) of a surface pixel area that have been averaged in order to reduce the effect of speckle, a noise-like consequence of the coherent nature of the signal transmitted from the sensor. The observed signals in the covariance representations, when multiplied by the number of looks, are complex Wishart distributed. This distribution is the multivariate complex analogue of the well-known chi squared distribution.

The complex Wishart distribution is completely determined by the parameters p (dimensionality), ENL, and Σ (the variance-covariance matrix). Given two observations of the same area at different times, one can set up a hypothesis test in order to decide whether or not a change has occurred between the two acquisitions. The null hypothesis, H_0 , is that $\Sigma_1 = \Sigma_2$, i.e., the two observations were sampled from the same distribution and no change has occurred, and the alternative (change) hypothesis, H_1 , is $\Sigma_1 \neq \Sigma_2$. Since the distributions are known, a likelihood ratio test can be formulated which allows one to decide to a desired degree of significance whether or not

to reject the null hypothesis. Acceptance or rejection is based on the test's p-value, which in turn may be derived from the (approximately known) distribution of the test statistic.

For analysis of the situation with data from two time points, $k = 2$, see [3, 4, 5, 6]. In [7] the authors describe bi-temporal region-based change detection for polarimetric SAR images by means of mixtures of Wishart distributions.

If we have data from more than two time points, $k > 2$, the procedure sketched can be generalized to test a hypothesis that all of the k pixels are characterized by the same Σ (the null hypothesis H_0),

$$H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_k (= \Sigma)$$

against the alternative (H_1) that at least one of the Σ_i , $i = 1, \dots, k$, is different, i.e., that at least one change has taken place.

For the logarithm of the omnibus likelihood ratio test statistic Q for testing H_0 against H_1 we have (see [1])

$$\ln Q = n\{pk \ln k + \sum_{i=1}^k \ln |\mathbf{X}_i| - k \ln |\mathbf{X}|\}.$$

Here n is ENL, the $\mathbf{X}_i = n\hat{\Sigma}_i$ (i.e., ENL times the observed covariance matrix) follow the complex Wishart distribution, $\mathbf{X}_i \sim W_C(p, n, \Sigma_i)$, and $\mathbf{X} = \sum_{i=1}^k \mathbf{X}_i \sim W_C(p, nk, \Sigma)$. Also, if the hypothesis is true ("under H_0 " in statistical parlance), $\hat{\Sigma} = \mathbf{X}/(kn)$. $Q \in [0, 1]$ with $Q = 1$ for equality.

The probability of finding a smaller value of $-2 \ln Q$ is approximated by ($z = -2 \ln q$, where q is the actually observed value of Q)

$$P\{-2 \ln Q \leq z\} \simeq P\{\chi^2((k-1)f) \leq z\};$$

$f = 9$ for quad pol, $f = 4$ for dual pol, $f = 2$ for dual pol diagonal only. The no-change probability is $1 - P\{\chi^2((k-1)f) \leq z\}$.

Furthermore this test can be factored into a sequence of tests involving hypotheses of the form $\Sigma_1 = \Sigma_2$ against $\Sigma_1 \neq \Sigma_2$, $\Sigma_1 = \Sigma_2 = \Sigma_3$ against $\Sigma_1 = \Sigma_2 \neq \Sigma_3$, and so forth. More specifically, to test whether the first $1 < j < k$ complex variance-covariance matrices Σ_i are equal, i.e., given that

$$\Sigma_1 = \Sigma_2 = \dots = \Sigma_{j-1}$$

then the likelihood ratio test statistic R_j for testing the hypothesis

$$H_{0,j} : \Sigma_j = \Sigma_1 \text{ against } H_{1,j} : \Sigma_j \neq \Sigma_1$$

is given by (see [1])

$$\begin{aligned} \ln R_j &= n\{p(j \ln j - (j-1) \ln(j-1)) \\ &\quad + (j-1) \ln \left| \sum_{i=1}^{j-1} \mathbf{X}_i \right| + \ln |\mathbf{X}_j| - j \ln \left| \sum_{i=1}^j \mathbf{X}_i \right|\}. \end{aligned}$$

Finally, the R_j constitute a factorization of Q such that $Q = \prod_{j=2}^k R_j$ or

$$\ln Q = \sum_{j=2}^k \ln R_j.$$

The probability of finding a smaller value of $-2 \ln R_j$ is approximated by $(z_j = -2 \ln r_j)$, where r_j is the actually observed value of R_j)

$$P\{-2 \ln R_j \leq z_j\} \simeq P\{\chi^2(f) \leq z_j\}.$$

The no-change probability is $1 - P\{\chi^2(f) \leq z_j\}$.

The tests are statistically independent under the null hypothesis. In the event of rejection of the null hypothesis at some point in the test sequence, the procedure is restarted from that point, so that multiple changes within the time series can be identified. For details also on better approximations to the distributions of Q and R_j under the null hypotheses, see [1, 8].

Since the omnibus method can detect not only if changes occur but also, within the temporal resolution of an image sequence, when they occur, long time series of frequent acquisitions over relevant sites are of special interest. One convenient source of such data is the Google Earth Engine¹ (GEE) [9] which ingests Sentinel-1 (and Sentinel-2) data as soon as they are made available by the European Space Agency (ESA) and provides an easy-to-use application programming interface (API) for accessing and processing the data.

3. CHANGE DETECTION AND RADIOMETRIC NORMALIZATION IN OPTICAL DATA

With respect to optical/visible-infrared (e.g., Sentinel-2 or Landsat) imagery, a data-driven, statistical approach to change detection is provided by the iteratively reweighted multivariate alteration detection (IR-MAD) algorithm [10, 4]. This method applies iterated canonical correlation analysis (CCA) to a multispectral images from two time points before performing band-wise differences. The CCA orders the image bands according to similarity (correlation), rather than spectral wavelength. The differences between corresponding pairs of canonical variates are termed the MAD variates. Specifically, a MAD variate Z is

$$Z = \mathbf{a}^T \mathbf{X} - \mathbf{b}^T \mathbf{Y}$$

where \mathbf{X} represents the m -dimensional image at time point 1, \mathbf{Y} represents the m -dimensional image at time point 2, and \mathbf{a} and \mathbf{b} are the eigenvectors from the CCA. Thus $\mathbf{a}^T \mathbf{X}$ is a canonical variate for time point 1 and $\mathbf{b}^T \mathbf{Y}$ is a canonical variate for time point 2. We have m uncorrelated canonical variates (CVs) with mean value zero and variance one from both time points, the correlation between corresponding pairs of CVs is ρ (termed the canonical correlation which is maximized in CCA), and we have m uncorrelated MAD variates with variance $2(1 - \rho)$.

In each iteration the values of each image pixel j are weighted by w_j which is the current estimate of the no-change probability and the image statistics (mean and covariance matrices) are re-sampled. Since the MAD variates for the no-change observations are approximately Gaussian and uncorrelated, the sum of their squared values (after normalization to unit variance)

$$C^2 = \sum_{i=1}^m \frac{Z_i^2}{2(1 - \rho_i)}$$

¹<https://earthengine.google.com> and <https://developers.google.com/earth-engine>

will ideally follow a chi squared distribution with m degrees of freedom, $C^2 \sim \chi^2(m)$. The probability of finding a smaller value of C^2 is approximated by (c^2 is the actually observed value of C^2)

$$P\{C^2 \leq c^2\} \simeq P\{\chi^2(m) \leq c^2\}.$$

Hence the no-change probability used as weight w_j in the iterations is $1 - P\{\chi^2(m) \leq c^2\}$. Iterations continue until the canonical correlations stop changing (or a maximum number of iterations is reached).

This procedure establishes an increasingly better background of no-change against which to detect significant change. Furthermore, canonical correlation analysis is invariant to linear and affine transformations, a fact that can be used to perform automatic relative radiometric normalization of the two multispectral images [11]. A threshold is set on the no-change probability (typically 95%) to identify invariant pixels in each scene. Their intensities are then regressed against each other band-wise to determine normalization coefficients. Because we have uncertainty in both variables here, we use orthogonal regression (as opposed to ordinary regression which places all uncertainty on the response variable). Again, the GEE is an ideal platform for accessing and processing (e.g., Sentinel-2 or Landsat) data in near real time.

4. CLOUD SOFTWARE

The authors have made available the necessary change detection software for interaction with the GEE on the open-source repository Github². The client-side programs run in a local Docker container serving a simple Flask web application. Apart from the Docker engine³ and a browser, no software installation is required whatsoever. After the user has been authenticated to the Earth Engine, he or she can carry out the following tasks: 1) run the IR-MAD algorithm on Sentinel-2 (or Landsat) bi-temporal imagery, 2) perform relative radiometric normalization in batch mode on an image sequence, 3) run the sequential omnibus algorithm on Sentinel-1 polarimetric image time series, 4) export imagery to his or her Earth Engine assets folder or to Google Drive for further processing or visualization.

(Software is available also for local processing. Tutorials on how to install software and to do both the polarimetric SAR and the optical data processing locally on your own hardware are available on Github.^{4,5})

5. EXAMPLES

To illustrate, the Sentinel-1 multi-temporal change map in Figure 1 displays the color-coded time intervals in which the most recent changes in the 2016 growth period in an agricultural area southwest of Winnipeg, Manitoba, Canada, occurred. The yellow and red areas (seasonally late changes) will mostly correspond to grain harvesting. The change maps can be viewed interactively in the GEE Code Editor.⁶

Figure 2 is a change frequency map showing shipping activity at the port of Tripoli, Libya, for a time series of 28 Sentinel-1 images. Heavy activity is concentrated to the northwest in the inner harbor.⁷

²<https://github.com/mortcanty/earthengine>

³<https://docs.docker.com>

⁴<https://mortcanty.github.io/src/tutorialsar.html>

⁵<https://mortcanty.github.io/src/tutorial.html>

⁶<https://code.earthengine.google.com/14d818dc83bed52608adf477999c76f8>

⁷<https://code.earthengine.google.com/5b543ad81805801d4c86a499bf4171a8>

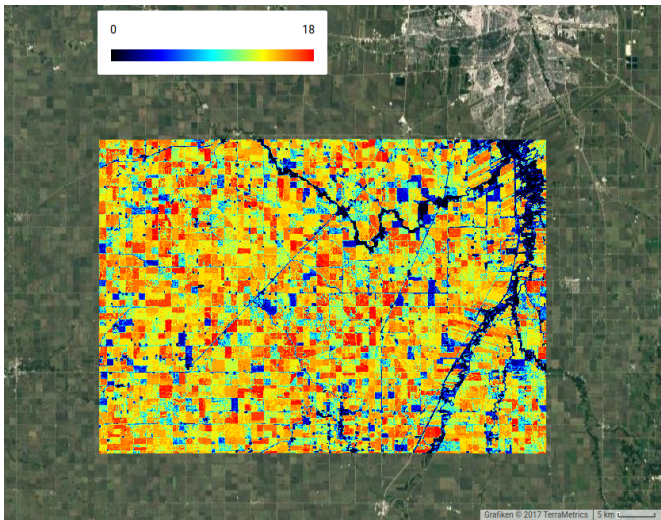


Fig. 1. Sequential omnibus change map for a region southwest of the city of Winnipeg, Manitoba, Canada, showing the time of the most recent change (black none, blue early, red late). The time series consisted of 19 Sentinel-1 images from May through October, 2016.

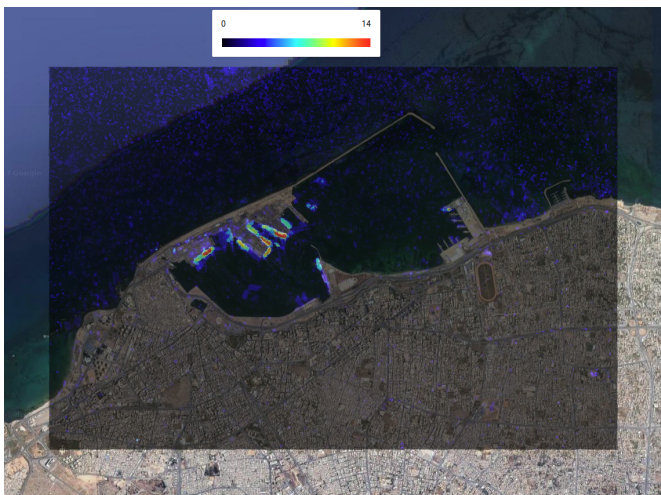


Fig. 2. Sequential omnibus change map for the port of Tripoli, Libya, showing the frequency of changes. The time series consisted of 28 Sentinel-1 images from April through December, 2016.

The golden yellow signal in the Sentinel-2 bi-temporal change map of Figure 3 shows part of the large area devastated by a major forest fire southeast of Coimbra, Portugal, which broke out on June 17, 2017. Note that the IR-MAD method clearly discriminates changes due to agriculture (in blue and cyan).⁸

The extreme flooding caused by hurricane Harvey in August, 2017 is apparent in the IR-MAD change map of Figure 4 (green signal).⁹ The heaviest rains fell between initial landfall near Houston, Texas, on August 26, continuing until August 29. We interpret the color graduation from green to blue at the edges of the flooding signal as reflecting receding floodwaters by August 30, the time of the sec-

⁸<https://code.earthengine.google.com/a1f9a4a55783c0e958941e56f150594c>

⁹<https://code.earthengine.google.com/b19e906e713448c862e512ccc8595b24>

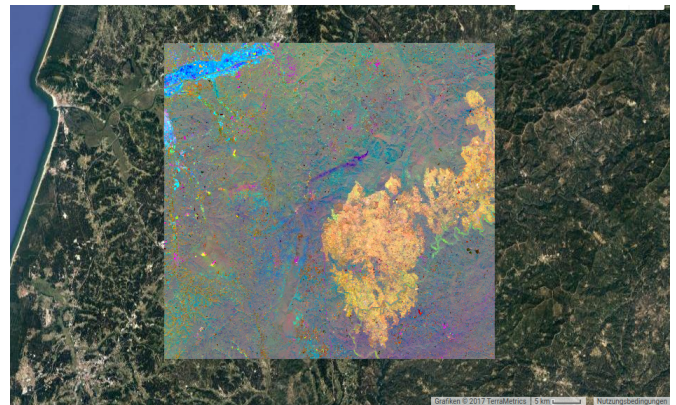


Fig. 3. IR-MAD bi-temporal change map (MAD variates 4, 3 and 1, where the variates are numbered from 1 to 4 according to decreasing canonical correlations as RGB) over an area southeast of Coimbra, Portugal, detecting a major forest fire. The two Sentinel-2 images used were acquired on April 4 and July 7, 2017. Only the 10m visual and near infrared bands 2, 3, 4 and 8 were processed.

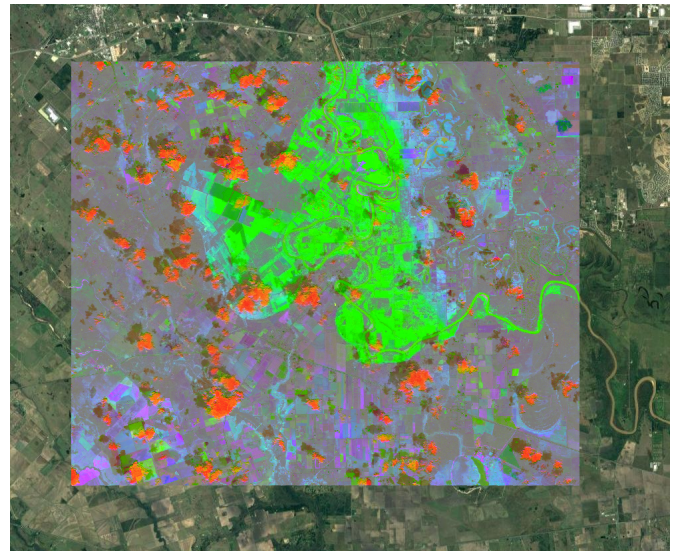


Fig. 4. IR-MAD bi-temporal change map (MAD variates 4, 3 and 2 as RGB) over an area west of Houston, Texas, USA, showing the flooding along the Brazos river due to hurricane Harvey. The two Sentinel-2 images used were acquired on August 20 and August 30, 2017. Only the 10m visual and near infrared bands 2, 3, 4 and 8 were processed.

ond acquisition. Note that the IR-MAD method clearly discriminates irrelevant changes due to cloud and cloud shadows (in red and dark gray).

Finally, Figure 5 illustrates relative radiometric normalization using two Landsat-7 ETM+ images.¹⁰ The first image (June 26, 2001) is used as reference, the second (August 29, 2001) as target, the target is normalized to the reference. Note, that the amount of change between the two acquisitions is considerable due to agricultural harvesting. Note also, that there is a clear difference in intensities espe-

¹⁰<https://code.earthengine.google.com/5f0c16f7922e9a7629971b7e393d00a8>

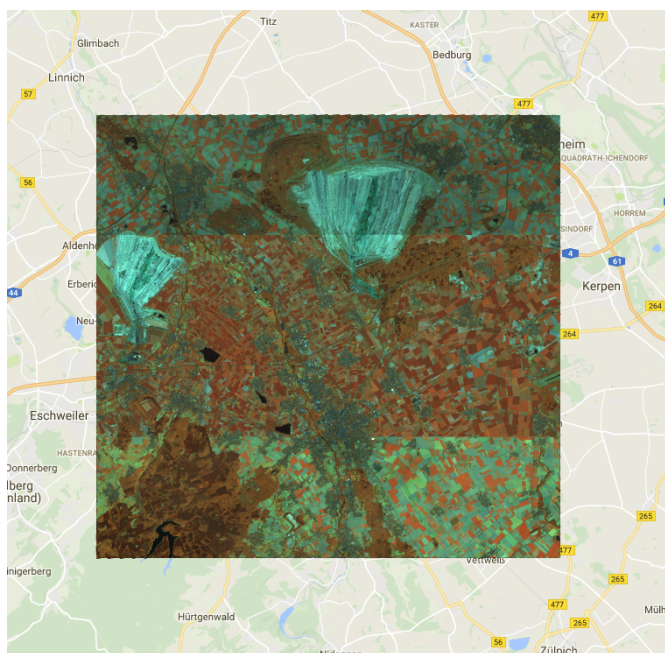


Fig. 5. Relative radiometric normalization of two Landsat-7 ETM+ images (at-sensor radiances expressed in digital numbers) acquired over the town of Jülich, Germany, on June 26 and August 29, 2001. Top segment: target image August 29, middle segment: reference image June 26, bottom segment: radiometrically normalized target image. Bands 4, 5 and 7 are shown in RGB composite linearly stretched from 0 to 250. The 30m non-thermal bands 1, 2, 3, 4, 5 and 7 were processed with the IR-MAD transformation to determine the invariant pixels.

cially noticeable in the open pit mine in the center of the transition between the original target and the reference (the top and middle segments) and that there, as desired, is no visible difference in intensities especially noticeable in the forested and urban areas in the left and center of the transition between the reference and the radiometrically normalized target (the middle and bottom segments).

6. CONCLUSIONS

Examples based on both Sentinel-1 dual polarimetry synthetic aperture radar data and Sentinel-2 optical data show the usefulness of the generic, automatic change detection techniques sketched. Note, that for the optical change detection method, because of the orthogonality between the change variates, different types of change can be discriminated between. Also, for optical data an automatic radiometric normalization scheme is sketched and illustrated. The examples shown cover different application areas: agriculture, surveillance/remote monitoring of port traffic and natural disasters, here forest fire and flooding.

Generic, automatic techniques as these are expected to be useful in many other application areas also where the study of spatio-temporal dynamics is important. The introduction of software available (to run either on your own hardware or) to anyone authenticated to run on the Google Earth Engine is expected to be extremely useful to researchers and practitioners alike.

7. REFERENCES

- [1] K. Conradsen, A. A. Nielsen, and H. Skriver, "Determining the points of change in time series of polarimetric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 3007–3024, 2016, Internet <https://doi.org/10.1109/TGRS.2015.2510160> and <http://www.imm.dtu.dk/pubdb/p.php?6825>.
- [2] J. J. van Zyl and F. T. Ulaby, "Scattering matrix representation for simple targets," in *Radar Polarimetry for Geoscience Applications*, F. T. Ulaby and C. Elachi, Eds. Artech, Norwood, MA, 1990.
- [3] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skriver, "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003, Internet <https://doi.org/10.1109/TGRS.2002.808066> and <http://www.imm.dtu.dk/pubdb/p.php?1219>.
- [4] M. J. Canty, *Image Analysis, Classification, and Change Detection in Remote Sensing, With Algorithms for ENVI/IDL and Python*, Taylor and Francis, Third revised edition, 2014.
- [5] A. A. Nielsen, K. Conradsen, and H. Skriver, "Change detection in full and dual polarization, single- and multi-frequency SAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 8, pp. 4041–4048, 2015, Internet <https://doi.org/10.1109/JSTARS.2015.2416434> and <http://www.imm.dtu.dk/pubdb/p.php?6827>.
- [6] V. Akbari, S. N. Anfinsen, A. P. Doulgeris, T. Eltoft, G. Moser, and S. B. Serpico, "Polarimetric SAR change detection with the complex Hotelling-Lawley trace statistic," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 3953–3966, 2016, Internet <https://doi.org/10.1109/10.1109/TGRS.2016.2532320>.
- [7] W. Yang, X. Yang, T. Yan, H. Song, and G.-S. Xia, "Region-Based Change Detection for Polarimetric SAR Images Using Wishart Mixture Models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6746–6756, 2016, Internet <https://doi.org/10.1109/TGRS.2016.2590145>.
- [8] A. A. Nielsen, K. Conradsen, H. Skriver, and M. J. Canty, "Visualization of and software for omnibus test based change detected in a time series of polarimetric SAR data," *Submitted*, 2017, <http://www.imm.dtu.dk/pubdb/p.php?6962>.
- [9] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Tau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017, Internet <https://doi.org/10.1016/j.rse.2017.06.031>.
- [10] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007, Internet <https://doi.org/10.1109/TIP.2006.888195> and <http://www.imm.dtu.dk/pubdb/p.php?4695>.
- [11] M. J. Canty and A. A. Nielsen, "Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1025–1036, 2008, Internet <http://www.imm.dtu.dk/pubdb/p.php?5362>.

OPERATIONAL APPLICATION OF THE FULL LANDSAT TIMESERIES TO SERVICE INDUSTRY IN THE AUSTRALIAN RANGELANDS

Peter Scarth

Joint Remote Sensing Research Program, School of Earth and Environmental Sciences, The University of Queensland, St Lucia 4072. p.scarth@uq.edu.au

ABSTRACT

Significant progress has been made in the development of cover data and derived products based on linking extensive field data to the full MODIS, Landsat and Sentinel-2 archives across Australia. These have led to the development of biophysical products tailored for the Australian Rangelands that are now used for quantifying and monitoring grazing land condition

These grazing land management products include fractional cover, fractional ground cover and persistent green state and trend products, as well as burnt area mapping. They also underpin several rangeland-specific information products used to assess the state and trends in rangeland environments; bare and green cover deciles to report on the current and historical condition of the grazing resource; and custom anomaly products to compare past and current conditions against a known baseline period.

To facilitate the interrogation and summarisation of these massive earth observation data sets in an accessible producer friendly way, a series of time series enabled web mapping and customised web processing services were developed, enabling the full time series over any spatial extent to be retrieved in seconds. These tools are being used by landholders monitoring paddock conditions, organisations supporting land management initiatives in the rangelands and Great Barrier Reef catchments, and researchers developing tools to understand land condition, degradation and human health across all of Australia.

Index Terms— Grazing, Landsat, Sentinel, Opendata, web services, timeseries

1. INTRODUCTION

The measurement of vegetation horizontal and vertical structure state and change are essential for mapping and reporting purposes in ecology, forestry, hydrology, agriculture and related areas. Data on vegetation structure are also used for state, national and international reporting, but over large areas are difficult and time-consuming to measure with the degree of precision required for monitoring purposes. Remote sensing can provide spatially- and temporally-comprehensive information about land cover features at a range of scales and often for minimal cost

compared to traditional mapping and monitoring approaches. This makes remote sensing a very useful operational mapping and monitoring tool for land managers, particularly across much of sparsely populated Australia.

Mapping and monitoring extent and change in groundcover and woody vegetation is a core requirement of several state and national government agencies to meet respective jurisdictions legislative requirements and to service a range of monitoring and reporting initiatives at local, state, national and global scales. Prior to open access of the USGS Landsat archive, this monitoring depended on sparse time-series of Landsat imagery to map wooded extent and Foliage Projective Cover [2]. Coordination of large area field observational databases [9] and open access to the full Landsat archive has opened new opportunities for advanced timeseries analysis and more frequent and rapid reporting on landcover change. The objective of this paper is to present an overview of recent developments undertaken by the Terrestrial Ecosystem Research Network (TERN) and partners to advance application of the full Landsat and Sentinel 2 archive to map and monitor condition and change in the Australian rangelands with a focus on the delivery of that information.

2. METHODOLOGY

In the Australian rangelands, TERN collaborates with several government and non-government organizations to help collect and organize field data as well as build scale and deliver biophysical products tailored to the monitoring of the Australian rangelands [1,7,8].

To support these products TERN partners have high-performance computing infrastructure to manage and process the Landsat archive for Australia. All available Landsat data is ingested into the system, process to surface reflectance [4], cloud and cloud shadow masked [5] and automatically processed into seasonal products representing three-month periods from 1987 until the current season [3]. This method of compositing has the benefit of reducing the frequency of spurious measurements, results in a regular temporal sequence, and allow seamless mosaics to be built at national scales.

The same processing chain is also applied to every Sentinel 2 tile across the east of Australia [5] to produce

seamless seasonal surface reflectance and fractional cover at 10m spatial resolution.

Fractional cover is a product with three values per pixel representing the fraction of bare ground, green vegetation and non-green vegetation. [10] used the field measurement database of [9] (Figure 1) to build a model which estimated fractional cover for each pixel in an image using a constrained non-negative least squares model. Due to the inclusion of training data from across Australia, the model provides reasonably accurate estimates (RMSE ~10%) of the cover fractions across a broad range of land types.

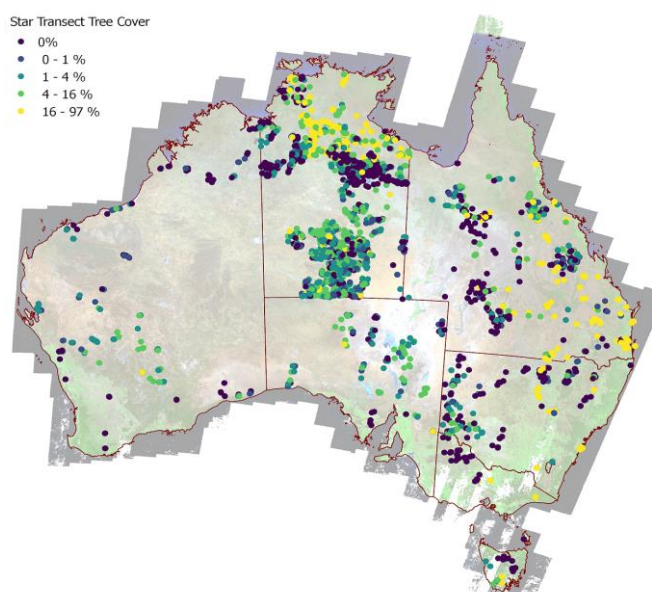


Figure 1 - Location of the 3000 star transect ground and tree cover plots across Australia. Each plot has 300 point intercepts within a 1 ha area and collects attributes about the amount of bare, green and non-green vegetation across the ground, mid story and over story components.

The resultant fractional cover product is used operationally for a range of applications across government and non-government organizations. The unmixing model can be equally applied to Landsat, Sentinel 2 and MODIS imagery as shown in figure 2.

The fractional cover product does not separate tree and mid-level woody foliage and branch cover from green and dry ground cover. Thus, in areas with even minimal tree cover (>10%), estimates of ground cover become uncertain. Therefore, we fit a spline to the base of the seasonal time series of green cover to estimate the amount of persistent green vegetation for any given pixel, providing an estimate of the woody vegetation. This separation of the woody tree and shrub layer, termed the 'persistent green', from the fractional cover product allows for the adjustment of the fractional cover image to create a fractional ground cover

estimate for each season, and these products are used by the grazing industry to inform on pasture condition and management effects. Other additional downstream products that are produced include the persistent green estimates, seasonal total and green cover deciles and dynamic reference cover products [1].

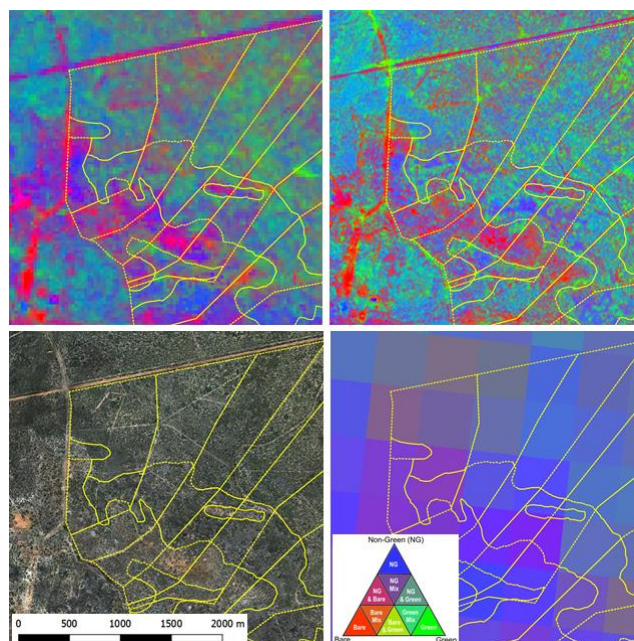


Figure 3 - Examples of seasonal fractional cover products over a grazing trial site in northern Australia. Clockwise from bottom left, true colour Worldview 3 image, Landsat 8 fractional cover (30m pixel), Sentinel 2 fractional cover (10m pixel) and MODIS fractional cover (500m pixel).

These biophysical information products are automatically uploaded as cloud optimized GeoTIFFs to open data portals¹ and are available by both direct download and through time enabled web services (Figure 4,5) that enable statistics on the complete 1988 to 2017 seasonal time series over any spatial extent to be computed and retrieved in seconds.

These services have enabled the development of additional downstream web processing services that allow the calculation of user-specified anomaly and grazing pressure analysis over individual properties and paddocks as well as time series-enabled web mapping and customized web-processing services. Since these services connect directly to the data on the portals, they can be set up as individual docker containers and deployed on infrastructure remote from the data without a significant loss of performance.

¹ <http://www.auscover.org.au/>

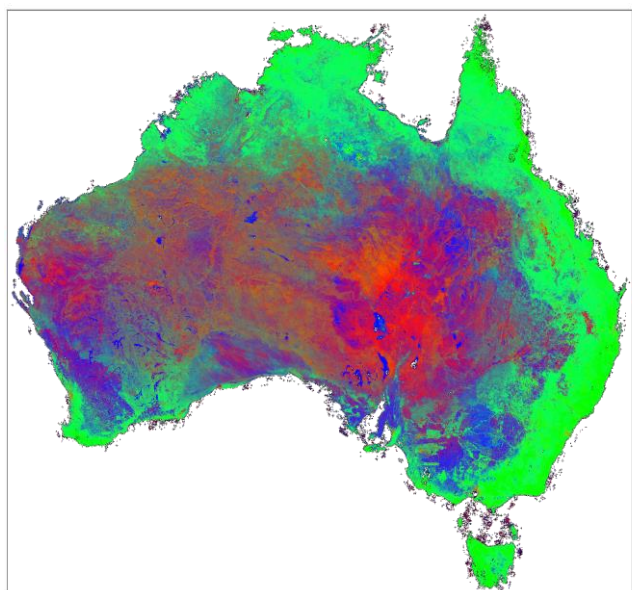


Figure 4 – Seasonal Landsat based fractional cover image, representing per-pixel proportions of green, non-green and bare cover visible from above. This product integrates the cover across all vegetation strata. Image is from January 2017 and accessed from time enabled web mapping services.

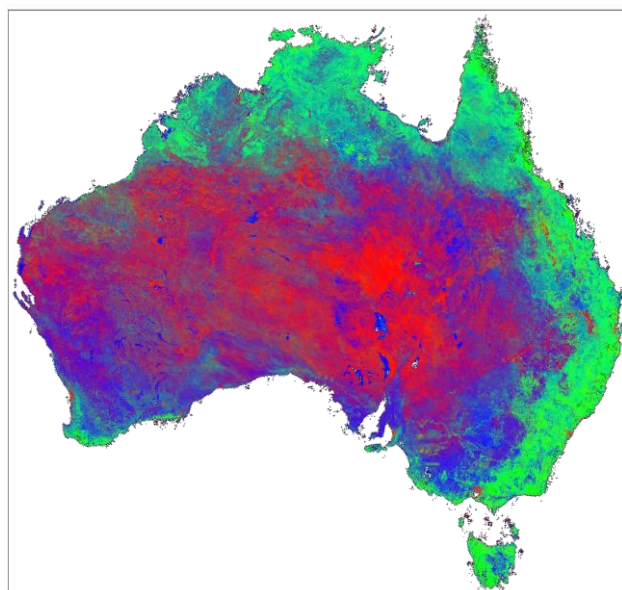


Figure 5 – Seasonal Landsat based ground cover image, representing per-pixel proportions of green, non-green and bare cover on the ground surface. The three and shrub signal has been removed. Image is from January 2017 and accessed from time enabled web mapping services.

3. RESULTS AND DISCUSSION

The production of seasonal composite from Landsat and Sentinel imagery and the delivery of these as cloud optimized GeoTIFF files has enabled a broad user base to easily access and use these products for analysis across many disciplines. By providing examples on how to access these data using off-the-shelf GIS software, python notebooks and API examples we are fostering uptake by cross disciplinary users to extend remote sensing science into their analyses and products.

The fractional cover and ground cover products along with derivatives including seasonal decile products, have now become embedded in many local state and national reporting frameworks due to their national coverage, widespread availability, and direct linkage to the field measurements across Australia.

In the grazing sphere, these tools allow interrogation and summarization of massive earth observation data sets in an accessible, producer-friendly way, and are being used by farmers monitoring paddock condition, organizations supporting land management initiatives in Great Barrier Reef catchments, and students developing tools to understand land condition and degradation and the underlying data.

There has also been significant commercial uptake with significant investment in producer focused web-based tools such as VegMachine² (Figure 6) and FarmMap4D³ which both connect to these data services. As of October 2017 we are seeing approximately 250 timeseries drills per day accessing the full 130 date seasonal archive through a time series API as well as a large number of web mapping tile deliveries as users browse, zoom and change dates to better understand the dynamic changes in their properties.

Feedback from producers, extension officers and developers are driving additional information products. The time series drill APIs now include the ability to provide climatic data including rainfall and accumulated rainfall sampled alongside the fractional cover information and these can be fed into server side and client side models to better understand the interactions between climate and management in these environments.

With the advent of higher spatial resolution data, such as that provided by the Copernicus Sentinel 2 series of satellites, we are starting to look beyond reporting purely on cover amount and more closely at the operational monitoring and reporting on spatial arrangement of cover and its links with land condition.

² <http://vegmachine.net/>

³ <http://www.farmmap4d.com.au/>

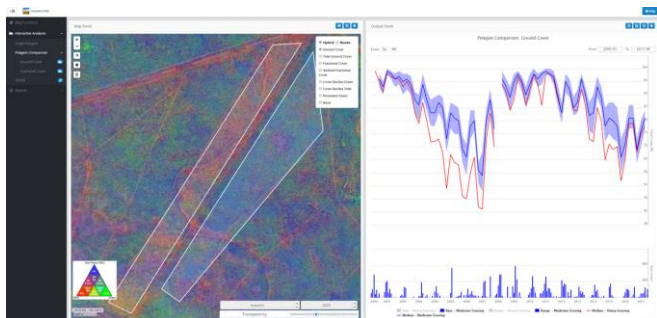


Figure 6 - Example property timeseries comparison calculated using the VegMachine website. Red line is the median ground cover timeseries for a heavily grazed paddock. Blue line is the median ground cover timeseries for a moderately stocked paddock and the transparent blue area represents the range of values seen in this paddock.

4 CONCLUDING REMARKS

By processing massive earth observation datasets into regular time series gridded biophysical products with integrated delivery methods, we can extend the use and integration of Earth observation across a much broader range of applications. These products and tools are seeing significant growth across Australia and this easy access to data is also driving development of new novel applications in cross disciplinary environments.

The metrics and products derived from this research our assisting land managers to prioritize investment and practice change strategies for long term sustainability and improved water quality, particularly in the Great Barrier Reef catchments. They are helping producers map monitor and understand the dynamics of their pasture resource over a more than 30 year timescale.

By relying on existing technology built into many open source tools we have been able to scale easily to national product generation and delivery as well as to incorporate new datasets such as Sentinel 2. Future work is now focused on building more tools on top of these services with a strong focus on delivering web time analytics to service real user needs across government and the public.

5. REFERENCES

- [1] Bastin, G. and Scarth, P. and Chewings, V. and Sparrow, A. and Denham, R. and Schmidt, M. and O'Reagain, P. and Shepherd, R. and Abbott, B. 2012. Separating grazing and rainfall effects at regional scale using remote sensing imagery: A dynamic reference-cover method. *Remote Sensing of Environment*, 121 . pp. 443-457.
- [2] Danaher, T., Scarth, P., Armston, J., Collet, L., Kitchen, J., and Gillingham, S. (2010). *Ecosystem Function in*

Savannas: Measurement and Modelling at Landscape to Global Scales. Vol. Section 3. Remote sensing of tree-grass systems: The Eastern Australian Woodlands. Taylor and Francis.

- [3] Flood, N., 2013. Seasonal Composite Landsat TM/ETM+ Images Using the Medoid (a Multi-Dimensional Median), *Remote Sensing*, 5: 6481-6500.

- [4] Flood, N., 2014. Continuity of Reflectance Data between Landsat-7 ETM+ and Landsat-8 OLI, for Both Top-of-Atmosphere and Surface Reflectance: A Study in the Australian Landscape. *Remote Sensing*, 6: 7952-7970.

- [5] Flood, N., 2017 Comparing Sentinel-2A and Landsat 7 and 8 using surface reflectance over Australia. *Remote Sensing*, 9 659.

- [6] Goodwin, N. R., Collett, L. J., Denham, R. J., Flood, N., and Tindall, D., 2013. Cloud and cloud shadow screening across Queensland, Australia: An automated method for Landsat TM/ETM+ time series. *Remote Sensing of Environment*. 134: 50-65.

- [7] Karfs, R.A. and Abbott, B.N. and Scarth, P.F. and Wallace, J.F. (2009) Land condition monitoring information for reef catchments: a new era. *Rangeland Journal*, 31 (1). pp. 69-86.

- [8] Metternicht, G., A Held, S. Phinn, R. Christensen, F. Kerblat, N. Sims, J. Guershman (2017) Earth Observation for supporting and tracking progress of sustainable development goals: best practice example from the Australian Terrestrial Ecosystem Research Network (TERN). 37th International Symposium on Remote Sensing of Environment, Tshwane, South Africa, May 2017.

- [9] Muir, J., Schmidt, M., Tindall, D., Trevithick, R., Scarth, P., Stewart, J., 2011. Guidelines for Field measurement of fractional ground cover: a technical handbook supporting the Australian collaborative land use and management program. Tech. rep., Queensland Department of Environment and Resource Management for the Australian Bureau of Agricultural and Resource Economics and Sciences, Canberra.

- [10] Scarth, P., Röder, A. and Schmidt, M., 2010. Tracking grazing pressure and climate interaction - the role of Landsat fractional cover in time series analysis. In: *Proceedings of the 15th Australasian Remote Sensing and Photogrammetry Conference (ARSPC)*, 13-17 September, Alice Springs, Australia. Alice Springs, NT.

PEPS – THE FRENCH COPERNICUS COLLABORATIVE GROUND SEGMENT

Stéphane Duprat¹, Driss El Maalem¹, Marc Ferrer¹, Vincent Garcia², Camille Louge¹, Mireille Paulin²
Erwann Poupart², Jérôme Gasperi,² Christophe Taillan²

1: Atos – 6, Impasse Alice Guy, 31024 Toulouse

2: CNES – 18 Avenue Edouard Belin, 31400 Toulouse

ABSTRACT

The aim of this paper is to present a quick overview of PEPS, the French “Sentinel collaborative ground segment by focusing on the main valuable services offered by the platform and to highlight the underlying technical solution.

As PEPS is part of a European Earth Observation ecosystem, this papers intends to highlight the relevance of this global environment.

Index Terms—Earth Observation, Sentinel, Big Data, PEPS, WPS, RESTo

1. INTRODUCTION

Europe’s investment in the Copernicus Sentinel satellites provides Europe with an unprecedented source of operational and valuable satellite data at global scale. Data streams are expected to amount to several terabytes per satellite orbit, thereby delivering unprecedented temporal and spatial resolution and data continuity.

This on-going global monitoring from space, combined with the heritage data and correlated with in-situ data, make up a unique and incredible amount of valuable information.

Nowadays, the improvement of the data quality (resolution, global coverage, ...), combined with an improvement of the revisit time related to a region of interest, makes it possible to think about new usage for the downstream & industry market as well as for science purpose.

Actually, all the conditions are met to open new data usages and create new opportunities that could not even be thought of some years ago by integrating Earth Observation data in the digital economy.

2. WHAT IS PEPS ?

PEPS - Plateforme d’Exploitation des Produits Sentinels - has been thought by CNES to facilitate the usage of Sentinel data by the user communities and thus boost the application development based on this tremendous space data & asset. Close to its users, listening to them to take into account their needs PEPS offers different access to data and processing capabilities. One of the main strength of PEPS is to minimize the publication



delays of the products after ESA production: 10% of the products are available in less than 3 hours and 80% of the products are available in less than 1 day.

On behalf of CNES, Atos is in charge from 2014 of the deployment of the required infrastructure and the development of the user services components and platform to reach out the user community.

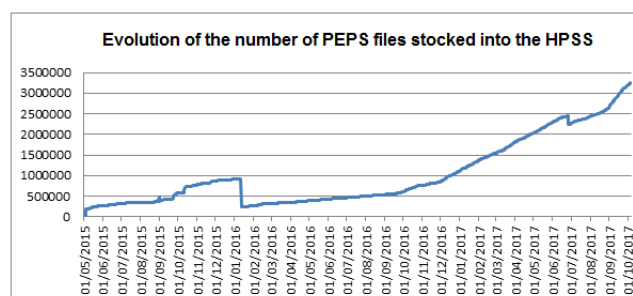
Starting from 2015, PEPS, as the French “Sentinel collaborative ground segment”, provides through its web portal - <http://peps.cnes.fr> – [1] a full access to all the Sentinel products according to the Copernicus, open and free, data policy:



Armed with a storage capacity of 14 Pb of data extensible up to 20Pb, PEPS provides the following services :

- On-line product user access
- Product catalog
- On-demand & automatic processing
- Discovery, view & downloading module
- Basic tools.

To this day, the PEPS archive contains 3,6 millions products and the PEPS website counts 2300 registered users.



3. TECHNICAL SOLUTION

The architectural solution has been designed to address the following technical challenges: modularity and scalability, performance, component reuse, communication by services, open source solutions and priority to the quality of service at user level.

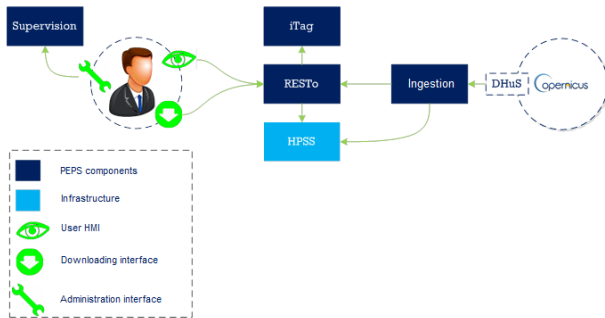


FIGURE 1 : PEPS GLOBAL ARCHITECTURE

3.1. Infrastructure

Software components are deployed on 3 servers completed by an HPC shared server

The ingestion server is connected to ESA's DHuS, the Data Hub Server for dissemination of the ESA Copernicus Sentinels data access (1.5 PB/month and more than 50000 users all over the world). The main function of this server is to regularly harvest sentinel data, store them on HPSS (High Performance Storage Systems), and communicate to the distribution server for catalog update.

Finally, the processing server and the HPC are in charge of the processing activities.

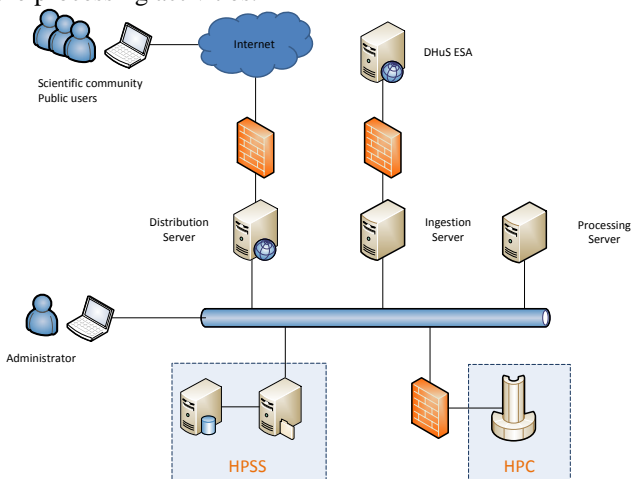


FIGURE 2: PEPS INFRASTRUCTURE

3.2. Architectural solutions

3.2.1. The ingestion module

Software components for the ingestion part are designed based on an **Actor Model** and implemented with Akka Scala. Motivations of that choice are scalability and Robustness.

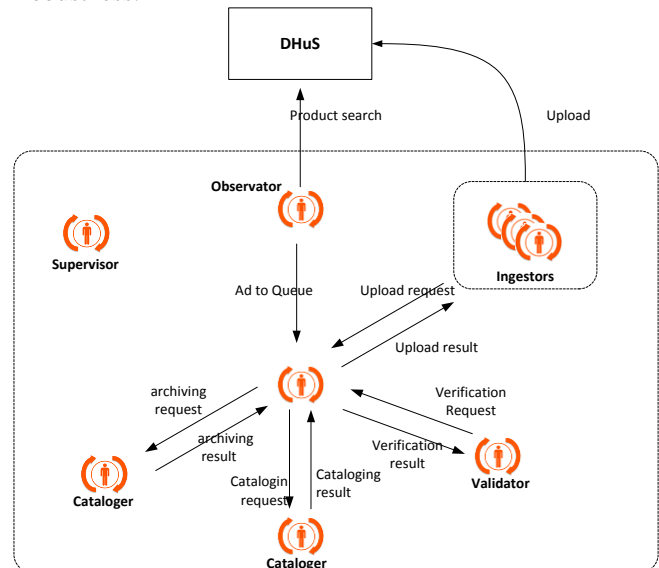


FIGURE 3 : INGESTION MODULE DESIGN

In this architecture, number of 'ingester' actors is dynamically adjusted relatively to each different data sources.

3.2.2. The cataloging and search module

Product **cataloging and search** engine are implemented by the RESTo [2] integration (Restful Semantic search Tool for geOSpatial) component. RESTo provides the following services: user management, cataloging, searching, cart management and downloading. All these services are provided through REST api allowing access from distant application.

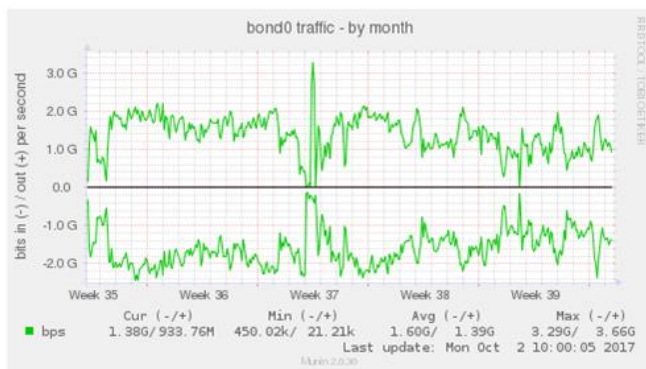
Based on these same services, a website offers a direct access to users.

RESTo implements different kinds of searches. It fulfills the **OpenSearch** standard with the extensions « GeoSpatial and Temporal » and « Extension for Earth Observation », thus enabling searches with spatial and temporal criteria. Users can also ask for a search expressed in natural language. The query is analyzed by the semantic analysis module of RESTo.

At the cataloging stage, PEPS also uses the iTag module which automatically tag geospatial metadata with geographical information (such as location, landuse, etc..)

3.2.3. Performance of the ingestion module

The ingestion rate reaches 1,6 Gbps in average.



3.2.4. The processing architecture

PEPS provides to users some ready-to-use **geospatial preprocessing services** following the WPS standard. These services allow users to remotely launch a processing task through the web interface or to call them directly through an API. The task can be executed either on the processing server or on the CNES's HPC shared server (High Performance Computing platform).

The Web Processing Service (WPS) Interface Standard provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services, such as polygon overlay.

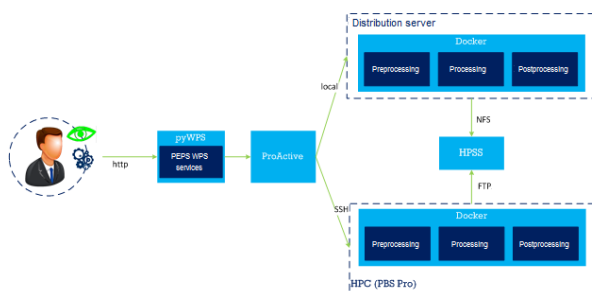


FIGURE 3: PEPS PROCESSING ARCHITECTURE

The PEPS processing architecture relies on the use of Docker.



Docker is a container technology for Linux that allows a developer to package up an application with all the parts it needs.



PEPS uses the ProActive [3] software to schedule and to orchestrate the processings on both the local distribution server and on the HPC.



As an implementation of the WPS standard, PEPS uses pyWPS [4] which is written in Python.

3.2.5. The available processings

PEPS provides different level of processing to different kind of users.

By the end of the year, all the registered users will have access to the first interactive tools through the website such as the quicklook generation of Sentinel 2 tile or the ortho-rectification of S1 imagery.

The CNES proposes to selected private companies on demand internal processing directly on the CNES infrastructure. The aim is to help those companies to test, tune and process closely to the data before migrating to an operational environment. To this date, 4 companies are concerned by this.

CLS -group : Determination of wind, wave, current, detection of ships, pollutions, detection of leads on sea ice ... on a global scale with Sentinel 1 data.



EXWEXs (Extrem Weather Expertises) associated to IFREMER: Integrated core forecast system with near real time Sentinel data.



ACRI-ST : Coastal change detection with Sentinel 2 data.

OceanDataLab : Full resolution display of multi-sensor data with Sentinel 1 and Sentinel 2 data.



3.2.6. The web client

Finally a web portal is provided to users providing access to data by search or by navigating on maps and to processings. It is developed with Angular JS technology and compliant with all web browsers.

The web client is responsive and is accessible through smartphones and tablets.

4. WHAT'S NEXT ?

PEPS guarantees to its users the transition to the integrated European platform *Integrated Ground Segment -Data and Information Access Service* (IGS-DIAS) which will be

provided in 2018. IGS-DIAS is an ESA initiative which aims to render Copernicus data and information available for access and further use in an efficient computing environment.

The service will continue in its present form up-to at least 2020. Currently on-going initiatives are focused on:

- The potential of « Web Processing Service » for on-line processing
- The integration flexibility of new algorithms in legacy processing chains (Continuous integration approach)
- The interoperability with others data & processing platform such as:
 - Data and Information Access Services
 - Thematic cluster such as Théia
 - Others national Sentinel collaborative ground segment such as CODE-DE
 - Thematic & Regional Exploitation Platforms

One of the main next stakes of PEPS is to provide the capability to interface with this valuable ecosystem in order to enable the design of processing chains and services distributed on various platforms.

5. REFERENCES

[1] PEPS Web site: <https://peps.cnes.fr/>

[2] Gasperi, J.: Semantic search within earth observation products database based on automatic tagging of image content. In: Proceedings of the Conference on Big Data from Space, pp. 4–6 (2014)

Resto : <https://github.com/jjrom/resto>

Itag : <https://github.com/jjrom/itag>

[3] ProActive Web site : <https://proactive.activeeon.com/>

[4] pyWPS Web site : <http://pywps.org/>

EUMETSAT, ECMWF & MERCATOR OCÉAN PARTNERS DIAS

Michael Schick, M.F. Dillmann, L. Wolf, J. Miguens, M. Stoicescu

EUMETSAT, Eumetsat-Allee 1, 64295 Darmstadt, Germany

ABSTRACT

EUMETSAT, ECMWF and Mercator Océan are jointly implementing a Data Information and Access Service (DIAS), which is centered on the “user to the data” paradigm. This undertaking has started in early 2017 and an initial Demonstrator version of this specific DIAS will be available in the middle of 2018. The paper is focusing on the Demonstrator.

Index Terms— Interoperability, Earth Observation, Big Data, Cloud Infrastructure, Copernicus Services

1. INTRODUCTION

The Integrated Ground Segment (IGS) Task Force was set up by the European Commission in July 2015 to establish a roadmap towards a more integrated Copernicus Ground Segment. Early 2016, with the support of this Task Force, the European Commission (EC) developed the concept of Data Information and Access Service (DIAS).

The primary purpose of DIAS is to ensure that the uptake of Copernicus data and information is maximized across a broad range of user groups and that Copernicus is able to provide the critical mass and focal point for stimulating innovation and the creation of new business models based on EO data and information, taking into account the challenges of “Big Data” and resolving the currently observed suboptimal situation of Copernicus data exploitation (see Fig. 1).

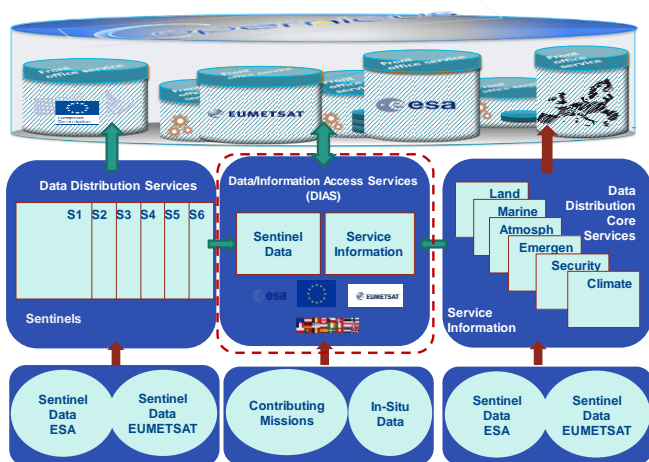


Fig. 1 Copernicus Eco-System

The Functional Requirements for the Copernicus Distribution Services and the Data and Information Access Services [1] established by the Commission and validated by the Copernicus governance provide with DIAS access to all Copernicus data and information in a more centralised and efficient manner (see **Error! Reference source not found.**). The DIAS include functions allowing users to discover, search and access the entirety of the Copernicus data and Information and offer also hosted processing capabilities available to third parties for developing and enabling Copernicus-related business opportunities.

EUMETSAT, ECMWF and Mercator Océan are bringing together their expertise to implement a DIAS. This endeavor builds on the EUMETSAT Data Services Roadmap which had already been initiated before the emergence of the DIAS concept.

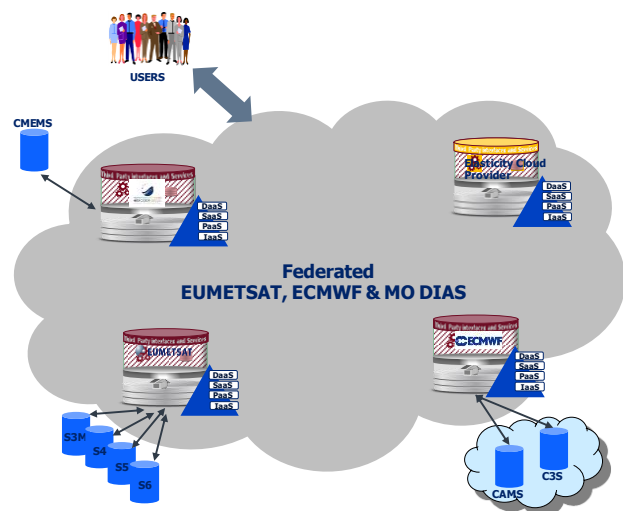


Fig. 2 EUMETSAT, ECMWF and Mercator Océan DIAS Concept

Conceptually, any DIAS instance consists of three building blocks: a back-office, an integration/interface layer and one or more front-offices. The back-office build on cloud computing principles will provide access to data and information (provided by the Sentinels, other contributing missions, in-situ data or data uploaded by users for their specific applications) and a scalable computing environment for its exploitation. The integration layer manages the interactions between front-office services and the back-

office. This layer consists of all interfaces that front offices require for their realization, while maximizing interoperability capabilities. The front-offices, either provided by the DIAS provider or developed by third parties (e.g. scientific institutions, SMEs, national meteorological offices), may provide value added services to their own user communities drawing on back-office capabilities.

2. USER SCENARIOS

The following user scenarios illustrate the functionalities of the DIAS demonstrator.

Users will be able to self-register and hence use after authentication DIAS demonstrator provided services. The service offer will range from a catalogue of all collections, harmonized data access to data and information, Virtual Machines (VM) and access to a Jupyter Notebook. Harmonised data access will support various filter criteria such as collection, time and area of interest.

Users are able to obtain access to Virtual Machines and use pre-installed tools (e.g. QGIS, SNAP) on the accessible data/information, as well as develop and execute their own applications by exploiting provided hosted processing capabilities. Instead of downloading TBs of data/information prior executing computations, users will be able to develop and execute their algorithms on the DIAS infrastructure, next to the data, and simply take away or share results of the computations.

Additionally, users shall be able to upload their own data, tailor their own development environments, benefit from user support and a service desk functions. Advanced users will be able to develop their own front-offices for their specific user communities.

3. DESIGN CHALLENGES AND OVERVIEW

Several mission exploitation platforms (MEPs) already exist, providing access to Copernicus data (e.g. EODC) or to other mission data (e.g. PROBA-V MEP). The novelty of DIAS lies in the amount and scale of the data volumes to be made accessible to the users (entire Sentinels archive and Copernicus Services Information) and in the data exploitation model articulated around front-offices. The success of DIAS will not be measured only in technical performances but also in the usage and the defined business model, including the market acceptance and user uptake.

In general the implementation of DIAS will be faced and addressing the following challenges:

- Diversity of User communities;
- Diversity of data and information;
- Disperse geographic locations (in a federation);

- Very large volumes of data;
- Interoperability and standardization of APIs;
- Business model (accounting & billing)
- Federated user management;
- Large scale resource usage & monitoring.

The DIAS functional requirements can be grouped into the functional blocks as shown within **Error! Reference source not found.** A description is provided in the following section describing the block coverage.

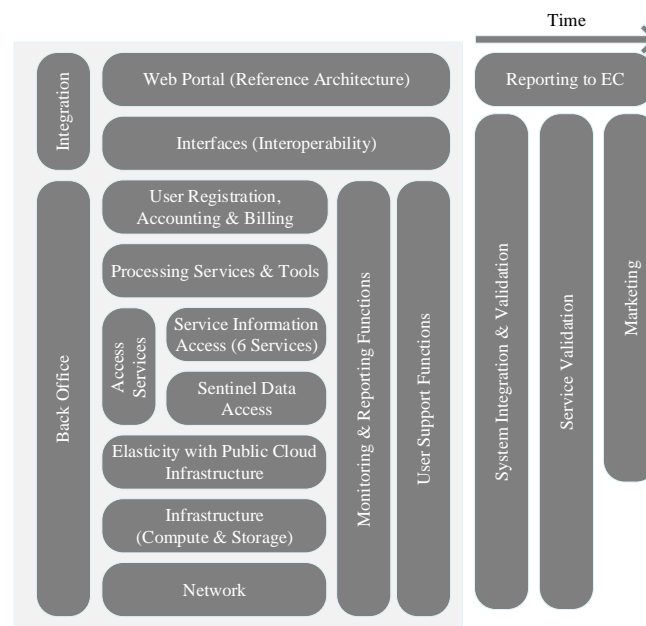


Fig. 3 EUMETSAT, ECMWF and Mercator Océan DIAS Concept

Web Portal (Reference Architecture) covers the web based front end to the DIAS. Its implementation will be based on the provided Interfaces block (APIs). The APIs will be used for example to: register users, interact with access services and list contents of the service registry.

User Registration, Accounting & Billing addresses all functionalities related to: user registration, user management, accounting, billing and license management.

Processing Services & Tools will allow end users to execute a “users to the data” scenario. Users will be able to use access services to exploit:

- Satellite data
- Service information
- Contributing mission data
- End users’ own uploaded data

Users will be able to interact with computing resources to invoke available software or their own uploaded applications interacting with harmonized access services. The processing includes functions to orchestrate workflows based on allowed quota and resources across time series of data and information. Pre-installed Tools (e.g. SNAP, Jupyter Notebook, QGIS), libraries and software development tools such as Python will be available to the end users for exploiting satellite data and information.

Access Services cover harmonised data access to all Data and Information based on an Adapter concept within the DIAS. This includes:

- Satellite data
- Service information
- Contributing Missions data

The management of licenses and copyright of data access will be covered including access rights and quota management. An overall collection & service registry will be managed, where users are able to browse through an overall inventory, prior accessing data or information.

Satellite Data Access provides harmonised access to satellite data via HTTP/REST interfaces. This functionality will build on Online Data Access services (e.g. EUMETSAT Data Services Roadmap – OLDA and also ESA DHuS).

Service Information Access will be managed within this functional block. The information collected here will be contributing to the service registry within the Access Services layer.

Elasticity with Public Cloud Infrastructure provides a mechanism for additional compute and storage resources for third party usage. Via this function, elasticity and on-demand scaling can be added for Third Party providers. This includes the required accounting and billing functionality such that a pay per use model can be put in place.

Infrastructure (Compute & Storage) functions are harmonised across the geographically distributed data centres, such as the distributed architecture is able to relocate the processing to as close as possible to the data.

Monitoring & Reporting Functions will take on all functions regarding the collection of monitoring information from all layers and its proper organisation, such that reporting and measuring of SLAs and fulfilment of KPIs can be checked. This will include dashboards/ graphs to efficiently work.

User Support Functions include the management of a service desk to support users. This includes maintaining a ticketing system and forum management services.

Interfaces and Interoperability functions are basically the API to the DIAS services. The role of this layer is to provide all capabilities required for a portal implementation to function. This is the case for the DIAS Web Portal. Elements such as service registries, exposure of provided APIs and interoperability arrangements belong within this functional block.

4. IMPLEMENTATION LOGIC

The full DIAS implementation is intended to follow a two-phase process. In the first phase (the demonstration phase), already started, the overall concept, high-level architectural design and deployment of a Demonstrator version, implementing the key aspects of the architecture and services, are performed. The Demonstrator version of the DIAS will be implemented in synergy with the on-going development of EUMETSAT's future Data Services roadmap, performed through EUMETSAT Multi Mission industrial service contracts. This demonstration phase is planned to be finalised during the first half of 2018 (see Fig. 4).

The second phase (the operational phase), dedicated to the operational implementation and subsequent operations of the DIAS in its full scope is intended to start in 2018, following major industrial procurement actions. These will consist of a set of new Copernicus specific competitive industrial service procurements for the full size implementation of the different system elements, integration, testing, operations preparation and operations of the DIAS, lasting until at least end of 2020. These major procurements will be initiated in the time period between end of 2017 and the first quarter 2018.

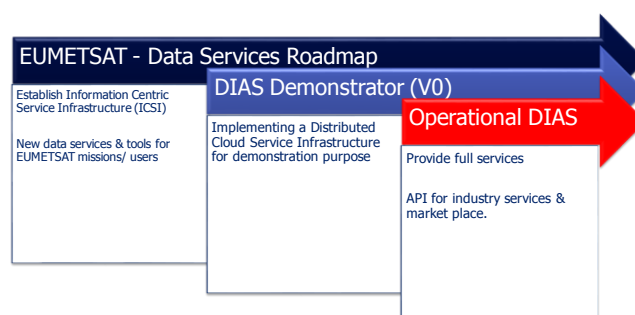


Fig. 4 DIAS development heritage

The EUMETSAT, Mercator-Océan and ECMWF DIAS concept and implementation are strongly driven by use cases and articulated around several key aspects, which make it attractive for implementation. It is from the start conceived to “bring users to the data” as a distributed system allowing efficient access to geographically distributed data and information. This feature integrates by design the aspect of interoperability with e.g. other instances of DIAS and national initiatives.

As a result, this DIAS is different to other implementation by its “federative” nature. Secondly, its development is incremental, starting with a pre-operational demonstrator of the key technologies, followed by regular releases implementing progressively the full scope of the DIAS and offering opportunity to correct/improve functionality based on regular received user feedback. Flexibility in capacity and speed is provided by cloud-based elasticity.

Furthermore, strong emphasis is put on the continued provision of highly reliable operational services with adequately manned user support functions, drawing on the expertise of the three partners. A fundamental aspect is that the three partners bring together know-how in complementary areas:

- EUMETSAT in providing 24x7 operations to end users, interoperability interfaces and data access & distribution
- ECMWF in high-performance computing and processing tools
- Mercator Océan in Web Portals, marketing and managing user communities.

5. CONCLUSION

This paper reflected on the current status of this specific DIAS and its early Demonstrator. Numerous technical challenges are to be faced and overcome for designing and developing this large scale, geographically distributed DIAS.

The challenges include, but are not limited to deploying an appropriate cloud computing environment at each partner’s premise, interconnecting them in a high-speed network and ensuring a homogeneous data access; not moving large volumes of data; setting up appropriate data caching mechanisms; fulfilling security requirements; ensuring user privacy and accountability; designing a sustainable and attractive billing model.

6. REFERENCES

[1] European Commission, “Functional Requirements for the Copernicus Distribution Services and the Data and Information Access Services (DIAS)”, <http://ec.europa.eu/DocsRoom/documents/20510/attachments/1/translations/en/renditions/pdf>

ASB – A PLATFORM AND APPLICATION AGNOSTIC SOLUTION FOR IMPLEMENTING COMPLEX PROCESSING CHAINS OVER GLOBALLY DISTRIBUTED PROCESSING AND DATA RESOURCES

Bernard Valentin¹, Matthieu Melcot¹, Leslie Gale¹

Philippe Mougnaud², Michele Iapaolo²

¹Space Applications Services NV/SA, Zaventem, Belgium

²European Space Agency / ESRIN, Frascati, Italy

ABSTRACT

The Automated Service Builder for Semantic Service Oriented Architecture project, ASB, had the challenge to create a common platform for the execution of heterogeneous processing chains (workflows) for processing facilities of Earth Observations missions with the aim to automate as much as possible the integration and provision of the processing facility services. A new application management paradigm based on the OGC® Web Processing Service (WPS) Interface Standard has been implemented together with task management that adapts automatically to the available ICT resources making it possible to tune ICT infrastructure costs to processing needs providing flexibility and dynamic scalability. Driven by the needs of the research and scientific community we realised that on-demand processing features could and should be added complementing the systematic processing capabilities. The result, a computing/data platform and application agnostic solution applicable to a wide range of processing needs and serving the research/scientific community and IT integrators needs over the full development cycle of processing facilities as well as Third Party Services developers. This paper presents what has been achieved and how it can be applied to creating facilities to process user defined workflows.

Index Terms — *EO data exploitation, Workflows, Processing chains, Processing facilities, Cloud Computing, Big Data*

Views expressed herein can in no way be taken to reflect the official opinion of the European Space Agency.

1. INTRODUCTION

The ASB project [1] was initiated in 2014 to investigate and demonstrate new approaches for developing the Processing Facility (PF) of a PDGS. In 2014 it was still common to have dedicated server based processing facilities. Software teams had the task of migrating or re-implementing the scientifically developed and tested algorithms from the prototyping phase, into operational systems taking into account processing and storage capacity limits imposed by the servers purchased by the mission for hosting the processing. The systems had limited flexibility and scalability. Cloud computing was seen to be an alternative, offering a solution with the potential to introduce common

services and because it is intrinsically elastic and scalable in principle removed concerns on the need to define a priori the processing and storage needs. Costs savings, development time scales reductions and flexibility to upgrade easily the processing facility with improved algorithms were expected advantages to be gained.

2. BACKGROUND

ASB is one project in a roadmap of activities being performed by the Space Applications Services Earth Observation Systems team investigating semantic technologies and the adoption of OGC standards. The focus of the roadmap is to create solutions to allow users to easily find and process EO data and products hosted on globally distributed resources.

In the time since ASB was initiated ESA/ESRIN has introduced Cloud computing for the delivery of their EO services and is investigating and supporting OGC standards in the context of exploitation platforms development (OGC Testbed13 EO Cloud [2]).

3. EVOLUTION OF THE USER COMMUNITY

For ESA EO instrument processing facilities (IPFs) the classical approach is to have a prototyping development phase where scientists (such as Principal Investigators, PIs) develop and document their algorithms followed by an implementation phase led by IT specialists who are tasked with optimizing and ensuring the algorithms are implemented to meet the requirements of an operational system. Although ESA has a Generic IPF Interface Specification [3] no obvious standard is available that would allow automated extraction of all of the information needed to create processing workflows and perform the orchestration of the processing chains. This knowledge is with the scientific team. It was clear by including and supporting the scientists all knowledge of the workflow could be captured offering the possibility for the scientists and the IT specialists to have a common picture of the workflows and provide a collaborative platform for the migration of the prototype developed algorithms to operational algorithms.

We recognize however that this is an initial step in the evolution of the user community wanting to implement complex processing chains. ASB makes it easy and attractive for "non-space" scientists, data analysts and

downstream service providers to access and use data processing facilities in complex workflows with an emphasis towards the services of Copernicus. ASB is designed to accommodate the evolution of the user community. The first step from facility developers to the science community was achieved offering a new operational mode of performing on-demand ad-hoc processing without disrupting the ASB core components.

4. ACHIEVED RESULTS

Space Applications Services has successfully completed the first phase of ASB innovatively combining existing technology with own developments to create a harmonised platform providing a generic, dynamic, scalable multi-mission (automated) processing environment.

An emphasis has been placed on being generic to create a mission (satellite and instruments), platform and application agnostic solution.

ASB makes it possible for users to define, configure and run algorithms embedded in workflows with an **Automated Generation of Workflows**. ASB provides functions to register new processes, graphically edit workflow definitions, executing processors with user-defined parameters, and access the results either through a product catalogue or an FTP server. ASB has implemented a customizable ontology-based mechanism that verifies the consistency of the dataflow between processes. Only compatible process parameters may be connected to pass data within workflows. Undefined parameters are treated as user defined inputs for which a Web-based interactive interface is automatically generated.

Generic Flexible Orchestration means that processor tasks are orchestrated by the workflow engine independent of the location of the actual executable files and independent of the underlying programming languages and related technologies. User algorithms are packaged, installed and executed on user-selected platforms which can be where the data is located.

Algorithm developers and software specialists benefit from **ASB support to Algorithm development** that allows user's own algorithms and libraries to be included in the ASB knowledge base as well as providing a platform to co-develop algorithms for faster development cycles, easy substitution of modules in processing chains for comparative studies and scalability at processing platform and algorithm levels.

5. ASB IMPLEMENTATION

5.1. Architecture

The ASB platform is built applying the microservices architecture (MSA) principles [4]. In order to facilitate both the installation and the distribution of the components, these

are deployed individually within Docker containers. Fig. 1 gives an overview of the platform architecture.

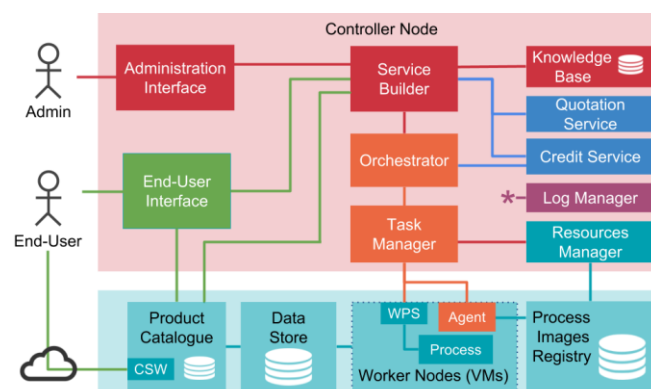


Fig. 1 ASB Architecture

There are two distinct environments: the Controller Node, and a Cloud-based Environment, in which Worker Nodes are deployed.

Controller Node – The ASB platform is made of a number of core components, exposing their functions through remotely callable (SOAP or REST) APIs making it possible to distribute the components on several (physical or virtual) hosts, should it be necessary. Together, the core components implement the Controller Node. Its physical location is of little importance. It can be deployed locally or remotely, in a dedicated server, or in a Cloud.

Core components include the Knowledge Base, which stores the definition of the data types, the processes and their input and output parameters, as well as the processors. The Service Builder is responsible for realizing the execution orders using the definitions and the user-provided inputs and for submitting these orders to the Orchestrator. The Orchestrator implements the workflow engine. It transmits individual process execution requests to the Task Manager. The Task Manager implements a queue and makes sure each process is deployed and run in an appropriate environment. The administration interface gives access to the platform internals. The end-user interface is used to request processor executions and access the results.

Cloud Environment and Worker Nodes – Worker Nodes are Virtual Machines (VMs) running in Cloud Environments. By default, these VMs run an Agent and have no pre-installed processes. Workflow processes are dynamically deployed and executed in the environment that meets their requirements. For example, a process that requires 16GB of RAM and a direct access to Sentinel-2 data will be deployed in an environment mirroring the Sen-2 archive and in a Worker Node that has at least 16GB of RAM available.

The actual workflow processes are stored in the Process Images (Docker) Registry and ready to be fetched and run by the Agents (see section 5.3, below).

5.2. Platform deployment modes

The platform has been developed in such a manner that the location and nature of the actual execution environment(s) are hidden from all but two of the core components: the Task Manager and the Resource Manager. All the other components are unaware of where and how the processes are executed. This makes it possible to deploy the software in a number of modes.

In **Stand-alone** mode where all the components as well as the processing chains are deployed and run on a single host. This setup is in particular convenient for developing and testing algorithms on small amounts of data and is particularly attractive to algorithm developers.

In **Cloud** mode, the process execution environment is a Private or Public Cloud, where resources are scalable. The available processing resources, in the form of VMs, are automatically detected and used to deploy processes. In this mode, intermediate and final products remain in the cloud, close to where they have been generated.

The **Cluster** mode is a restricted Cloud-mode in which the available processing resources are considered as fixed, with no possibility to scale up or down the cluster.

5.3. Dynamic behaviour

The collaboration diagram shown in Fig. 2 depicts the components and the interactions that take place at the time a process execution request is received by the Task Manager.

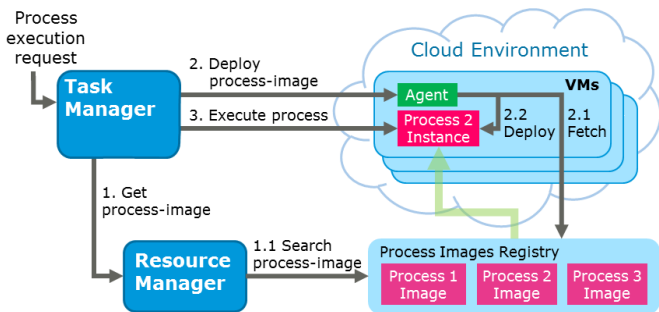


Fig. 2 Dynamic Process Deployment

The following interactions take place when a process execution request, issued by the Orchestrator (Workflow Engine) component, is received by the Task Manager:

1. The Task Manager communicates with the Resource Manager to determine in which Process Images Registry the required process image is available.
2. The Task Manager selects a Worker Node having the resources (CPU, memory, etc.) required to run the process and asks the node Agent to fetch and instantiate the process image. If no new instance may be deployed (lack of resources in the node), the execution request is queued until a suitable node becomes available.
3. As soon as a process instance becomes available, the Task Manager executes the process and polls the

service until it returns a result. The process outputs are then recorded and provided back to the requester.

6. USING ASB

The following sections briefly describe the use of the ASB platform, as a developer and as a user, and introduce the projects in which ASB is currently used.

6.1. Preparation of a Custom Processor

The creation of a custom processor is performed in two steps. In this procedure, the user gives the platform all the elements required to successfully execute the processor:

1. All the processes to be executed by the processor must be deployed and registered in the platform. If no new process is required, this step may be skipped.

In the current version of the platform, the deployment and registration of new processes must be performed manually. A deployment tool is in development that will automate most of this work.

The manual deployment of a process consists in (1) creating a Docker image, using a parent image that contains a pre-deployed WPS [5] server and other dependencies, (2) pushing this image in the Process Image Registry, and (3) registering the new image in the Resource Manager. The process metadata, including the definition of its inputs and outputs, must then be registered in the Knowledge Base.

2. The new processor is defined via the Knowledge Base Web interface. This includes a form for entering the processor properties, and a graphical workflow editor (Fig. 3) for identifying the processes and specifying the flow of the data between these processes (output to inputs connections). The Processor Workflow Editor displays the list of available processes, an interactive graphical workflow editor, the short description of the selected process, and a preview of the user parameterisation form. This form is dynamically generated and includes all the process inputs present in the workflow not connected to a process output.

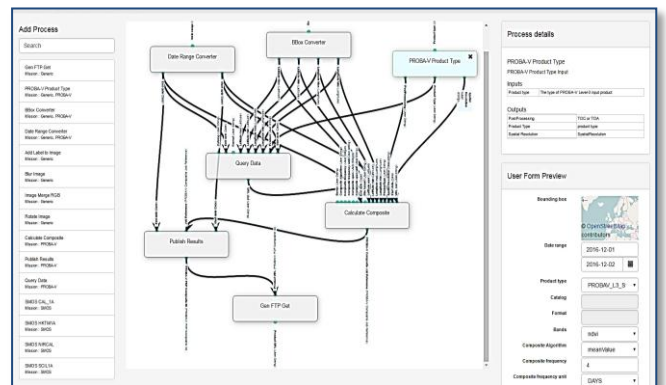


Fig. 3 Processor Workflow Editor

Configured processors are displayed in the user interface. They may be inspected and selected for execution, as described below.

6.2. Processor Execution

Users select the processor to be executed in the user interface, access the dynamically generated parameterisation form (Fig. 4), fills-in the required inputs, and triggers the execution. Progress of the execution is available by means of a percentage of executed processes. When the execution is complete, a report is generated which contains information about the execution as well as links to the generated products.

Fig. 4 Parameterisation Form

6.3. Examples and Lessons Learnt

Two use cases were successfully demonstrated in the course of ESA's ASB project. Case 1: systematic processing. For part of a SMOS processing chain executables, as-is from the SMOS IPF, were integrated in a workflow to manage and control auxiliary input file dependency. Case 2: processing on-demand. Using an automatically generated form users provide inputs to compute PROBA-V radiometric and NDVI composite products.

ASB is now being used and improved within ESA project PROBA-V MEP Third-Party Services (TPS [6]) for supporting scientists. In particular, the Desert Locust Habitat Monitoring application developed by the *Université catholique de Louvain* (UCL) is deployed and run with ASB within a cluster hosted in the PROBA-V MEP.

ASB has permitted to encapsulate algorithms written by scientists within process images, deploy these images in the environment, configure the application processor via the user interface, and execute the application on NRT data with a highly reduced effort, and more particularly no specific coding. The result is a processor that runs the original algorithms, tested and validated by the scientists, in an operational context. The ASB components being generic and processing environment agnostic, this scenario may be easily re-applied to other applications and missions.

A proof of concept is being implemented for performing globally change detection using Sentinel-1 data for crisis management requiring rapid scalability.

7. FUTURE

ASB assumes that technology will continue to develop and that the lack of standardisation of interfaces and mechanisms such as the packaging and running workflows in the cloud would eventually be tackled. In OGC Testbed 13 EO Cloud [2] a first step to define standards for packaging and running workflows has been taken but it does not fully address aspects required by ASB such as the ability to directly deploy the algorithms written by the users, instead of requiring the users to do the packaging (Docker containerization) themselves, and the ability to configure and run workflows for orchestrating the deployed services. This will be addressed in H2020 Big Data Shift project EOPEN [7] in which Space Applications Services is leading the implementation of the processing platform. Part of the work will be to extend the platform for supporting the deployment of processes in heterogeneous environments including Clouds and HPCs.

8. CONCLUSION

The ASB project has shown that a generic platform can provide a dynamic, scalable and automated processing environment to execute processing chains triggered automatically when new data becomes available or on-demand via a graphical user interface or through a WPS processing service.

ASB is being used in projects and is available [8] and ready to provide an environment allowing users to access and use multiple platforms aggregating the capabilities of globally available data and processing resources such as DIAS and other exploitation platforms.

9. REFERENCES

- [1] ASB project page on ESA Research & Service Support, <https://wiki.services.esa.int/wiki/index.php?page=ASB>
- [2] OGC Testbed13 EO Cloud, <http://www.openeospatial.org/projects/initiatives/testbed13>
- [3] "Generic IPF Interface Specifications," ICD MMFI-GSEG-EOPG-TN-07-0003, ESA, August 2009.
- [4] "Microservices: a definition of this new architectural term", James Lewis, Martin Fowler, 25 March 2014, <http://martinfowler.com/articles/microservices.html>
- [5] "Web processing service 1.0.0," OGC 05-007r7, OGC, June 2007. Corrigendum, OGC 08-091r6, June 2009.
- [6] PROBA-V MEP Third-Party Services (TPS) project, <https://proba-v-mep.esa.int/proba-v-mep-tps-users>
- [7] EOPEN, H2020 project, opEn interOperable Platform for unified access and analysis of Earth observation data
- [8] "Information on becoming a user of ASB," <https://asb.spaceapplications.com/demo>

ONEATLAS
AIRBUS DEFENCE AND SPACE DIGITAL PLATFORM FOR IMAGERY
Laurent Gabet, Philippe Nonin, Salvador Cavadini, Mathias Ortner, Mathieu Rouget
Airbus Defence and Space

ABSTRACT

Airbus Defence and Space has developed an online platform providing clients with access to a global satellite image data base map at 1.5m and 0.5m resolution, created and updated automatically with data collected each day by satellites managed by Airbus. The service was designed and deployed in a fully scalable public Cloud infrastructure to maximize efficiency and to allow development of additional analytical services. The platform development automated existing processing flow lines enabling new satellite data to be uploaded onto the Cloud platform at a rate of 1 new image every 3 minutes, 24hrs a day. Whilst technically innovative this service also enabled the transformation from a traditional market pull business model to a new market push business model, i.e., supplying data via annual subscription services. It also forms the basis for offering a new generation of online information and analytical services, directly to customers via API's.

Index Terms— EO satellite images, Analytics, Platform, Cloud, Massive Processing

1. INTRODUCTION

Earth observation satellite imagery landscape is drastically changing. Customers are more and more seeing satellite imagery as a commodity for achieving their business goals. Accessing satellite imagery should not be an issue any more. Customers' expectations are to access images in real time with no delay, the freshest together with the older. Accessing an image does not mean anymore to download and process the image in its own environment. It means dealing with the images in real time from the office but also from online platforms. Images are transformed; features are extracted to complement Big Data analysis, providing meaningful information.

Airbus Defence and Space is using an EO satellites constellation of 6 flag ships, four optical and two radars. Four new optical satellites (30cm GSD) to be launched in 2020 are under construction. The new constellation will have an acquisition productivity far larger than the current one. All together multiple petas of images will have to be stored each year.

To be in line with the market evolution and the massive amount of data to manage, Airbus Defence and Space started early 2015 the development of a cloud based digital platform specialized on EO images. The first version of the platform has been commercially opened mid-2016. Developments are still on going to add new analytics functionalities.

The article is describing the key challenges which have been faced and technical features implemented to setup a very attractive and efficient solution. The platform has been developed on the newest cloud technologies and is addressing in a secure mode the new EO challenges.

OneAtlas is the commercial name of Airbus Defence and Space digital platform for EO imagery. It covers the storage (on line image archive), the processing functionalities, the services and the analytics.

2. PLATFORM DRIVERS

The platform has been implemented to fulfill a large set of constrainable requirements:

- All images should be accessible in real time, not only the freshest one but also all the historical one.
- Tens of Petas of data should be stored in a cost effective way with no compromise on the accessibility.
- Streaming capacities should be offered for professional use, meaning that visualization time should be less than 1 second.
- Only images at the lowest processing level (raw) should be stored, all other levels, ortho by example, are processed when required.
- Real time visualization of orthoimages and mosaic should be possible without preprocessing them.
- Elastic scalable capacities should be provided: no limits on massive processing and on volume of images.
- No reduction of the processing efficiency should occur during massively parallel processing.
- All the different types of analytics / processing should be supported: Standard photogrammetric (including 3D), remote sensing processing but also Machine Learning.

- Access to the platform should be done through APIs to enhance platform to platform capacities and automatisms.
- Should be based on a public cloud infrastructure to enable the scalability, reduce the investment and provide internet fast access. But with no compromise on the security.
- Partners should be able to add their own processing tools

3. ARCHITECTURE

In Cloud infrastructures, processing (compute nodes) and storage (object storage) are separated. Historical EO processing has been developed in a POSIX environment meaning that the compute nodes access to the images through a file system (physical disk, SAN, NAS, nfs...).

To deploy EO processing on the cloud, most of the current platforms have implemented an intermediate block storage space in between the compute nodes and the object storage. It gives to the compute nodes an access to the images in a standard POSIX way. This approach provides the advantage to be able to migrate with minimal efforts the existing processing algorithm to the cloud. But it requires either to store all the images in the POSIX file system or to copy them from the object storage to the POSIX disk when required. This approach has a major limitation in terms of storage scalability. It is also a very costly solution which is not sustainable for a very large amount of images.

Airbus Defence and Space designed an innovative platform architecture to solve this issue. The innovation is very simple in the concept; it connects directly the compute nodes to the object storage (Figure 1). No intermediate POSIX file system is required. The direct connection is provided through a driver (called DAAS). The driver provides simple GDAL access functions: get and write buffer.

To make the driver efficient it has been necessary to design an innovative image container. The images are containerized before being stored in the object storage. The container acts as a facilitator between the driver and the object storage.

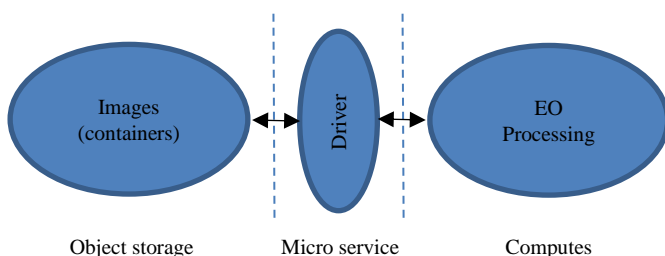


FIGURE 1: CLOUD PLATFORM MACRO ARCHITECTURE

This container takes inspiration from existing image formats (GeoTIFF [1]) and big data formats (Avro [2], Parquet [3]). It is a sparse collection of multi-layer tiles and is versatile enough to represent very large images, such as continent-scale high resolution EO products.

When designing this image container, we followed two main drivers. First, we aimed at lowering the operating costs on both private and public clouds. The container supports several compression algorithms and implements vectorized reads and writes operation that reduce IOs to the Object Storage. Second, the container offers the optimal performance for both low latency processing and massive batch processing operations. It is fully scalable and support massively parallel and distributed read and write operations on Object Storage (Figure 2).

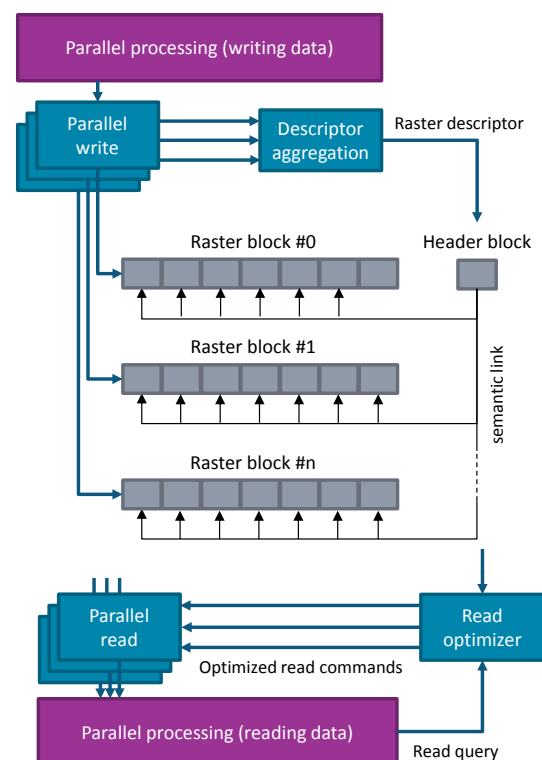


FIGURE 2: IMAGE CONTAINER READ AND WRITE OPERATIONS

4. PROCESSING FRAMEWORK

The processing framework is implemented following a microservices-oriented architecture style; that is: the framework is structured as a collection of loosely coupled services. That kind of organization promotes modularization of the application facilitating development, testing and comprehension of the system as a whole and its composing

parts. Moreover, microservices-oriented architectures are well suited for addressing main concerns of the platform design drivers: elasticity, platform variety, and performance constraints.

The choice of microservices architecture style implies the need of a framework able to deploy, manage and monitor microservices in such a way that elasticity and scalability are facilitated. The processing framework relies on Docker (<https://www.docker.com/>) and Kubernetes (<https://kubernetes.io/>) for packaging, automating deployment, scaling, and management of microservices.

The processing framework exposes its functionalities through RESTful APIs. Two kinds of geo-processes are exposed:

- On-the-Fly (near real time) processes, and
- Massive (batch) processes

On-the-Fly (OTF) processes are implemented as synchronous services that, for the sake of performance, minimize interactions with other services. Due to the near-real time nature of the processing error recovery strategies are not implemented therefore as soon an error occur the process fails and returns.

Two main pipelines have been optimized following the OTF approach: ortho rectification and ortho mosaic.

On-the-fly processing is evidently not applicable when processing large amounts of data. That is because massive processes may take times to completion by orders of magnitude bigger (e.g. hours) than OTF processes. OTF processes are implemented as synchronous calls between the requester and the process implementation; synchronous implementation of long-lived processes requires maintaining a connection between the caller and the callee. If the connection is closed, due to communication layer errors or failure of any of both sides, the process fails and all progress made until the closing of the connection is lost.

The processing framework provides tools for implementing massive processes (a.k.a. processing pipelines) by composing services.

The framework provides a set of orchestration ([https://en.wikipedia.org/wiki/Orchestration_\(computing\)\)](https://en.wikipedia.org/wiki/Orchestration_(computing))) *primitives* and *concepts* to build processing pipelines. While primitives are implemented as independent and autonomous services, concepts allows to take care of the flow of control and data in pipelines. Orchestration primitives are exposed as microservices. Geo-processing pipelines are created by composing orchestration primitives and geo-processing micro-services.

Routing data through pipeline's processes, pipeline *dataflow*, is achieved by using a combination of two approaches: *direct flow* and *indirect flow*.

In *direct flows*, data produced by a process is routed, without intermediates, to the process that consumes it. Direct flows are the most efficient mechanism to pass data between processes but it is not always possible or convenient to use them; in such cases data flows can be implemented by using *indirect flows* where data produced by a process is put in an *intermediate storage* to be accessed by other processes later.

Indirect flows allow overcoming the limitations of direct flows but at the price of performance and complexity overhead.

In order to overcome input data latency in on-the-fly applications, processing schemes have thought as pure pipe line, invariant to input random order of arrival.

The asynchronous nature of massive processes requires the implementation of *process monitoring* functionalities. Users need to access information on the status of their processes. Massive processing framework provides such functionalities through the *Monitoring service*.

Monitoring service is able to:

- Respond to user requests about processes status
- Notify users about processes status

To accomplish his task, Monitoring service uses the data from the *Events store*, a database that collects all events related to processes executions.

In the context of a pipeline execution, orchestration primitives send *execution events* to the Events store. Events are, basically, information related with the start and the end of processes executions. Using these events, the Monitoring Service is able to calculate the current status of processes and inform the user.

To insure an efficient horizontal scalability and fault tolerant workflows within the cloud, all the photogrammetric and remote sensing algorithms have been split in functional modules designed according to the following guidelines:

- Stateless processes.
- No side effect (pure functions) i.e., the service has no observable interaction with its environment other than returning a value.
- The task granularity should be set as the maximum admissible resource loss in case of failure.
- Agnostic of framework primitives. That is, processing services cannot deliberately use primitives of the framework.

5. ANALYTICS

A set of analytic services have been deployed on the platform.

Boat detection is an example of an analytics using Machine learning. A Tensorflow[4] model has been trained in a development environment on the cloud (Figure 3). All the images available on the OneAtlas platform may be used for the training phase.



FIGURE 3: BOAT MACHINE LEARNING CLASSIFIER USING TENSORFLOW

At the end of the training loop the Tensorflow model is encapsulated in a docker (Figure 4) and exposed as a service in the platform.



FIGURE 4: INGESTION OF THE TENSORFLOW MODEL IN THE PLATFORM

The service API can be used to extract the boat (Figure 5) from any image available in OneAtlas on-line image archive.



FIGURE 5: MACHINE LEARNING BOAT EXTRACTION FROM A SPOT 6 IMAGE

6. CONCLUSION

OneAtlas platform specifically designed for image processing on the cloud has been introduced. Key features to make it efficient and sustainable have been described: direct connection between object storage and processing modules, refactoring of the code to adjust to cloud constraints, processing frame work. An example, boats detection, of application running on the platform has been introduced.

The next phases are: finalizing the industrialization of the platform (PAAS, SAAS) and welcoming more processing.

7. REFERENCES

- [1] <http://trac.osgeo.org/geotiff/> – GeoTIFF image format specification
- [2] <http://avro.apache.org/docs/1.8.2/spec.html> – Apache Avro 1.8.2 specification.
- [3] <https://parquet.apache.org/documentation/latest/> – Apache Parquet design
- [4] Google Research, “Tensorflow: large-scale machine learning on heterogeneous distributed systems”, White Paper, 2015

PROBA-V MISSION EXPLOITATION PLATFORM

Erwin Goor, Jeroen Dries, Dirk Daems

VITO, Boeretang 200, 2400 Mol, Belgium, <http://www.vito.be>

ABSTRACT

VITO and partners developed and currently operate an end-to-end solution to drastically improve the exploitation of the Proba-V and SPOT-VEGETATION EO-data archives [1] and derived vegetation parameters from the Copernicus Global Land Service [2] by researchers, service providers and thematic users. The analysis of time series of data (+1PByte) is addressed, as well as large scale on-demand processing of the complete archive, including near real-time data.

Since January 2016 a pre-release of the platform was available, as presented at the BIDS'16 conference, and since November 2016 the **Proba-V Mission Exploitation Platform (MEP)** is released as a fully **operational service** to our users. Several applications are provided, e.g. a time series viewer, a full resolution GEO viewer, pre-defined on-demand processing chains and virtual machines with powerful tools and access to the full data archive. This allows users to design, debug and test applications on the platform. All these services are accessible from <https://proba-v-mep.esa.int> [3]. In the next two years, the platform will involve significantly and the user support is intensified. Furthermore access to e.g. Landsat-7/8 and Sentinel-2/3 data on the platform is being addressed as well.

Index Terms— MEP Mission Exploitation Platform, Proba-V, vegetation, data analytics on time series, on-demand processing, virtual research environment

1. OBJECTIVES AND BENEFITS

The Proba-V MEP complements the Proba-V user segment by building an operational Exploitation Platform (EP) on these data, complementary data and derived products, addressing hereby the wider vegetation user community with the final aim to ease and increase the use of Proba-V data. The data offering consists of the complete archive from SPOT-VEGETATION, Proba-V and bio-geophysical parameters from the Copernicus Global Land Service.

The reasons for deploying a MEP dedicated to the Proba-V mission are numerous:

- The data and specifically the time series of daily / ten-daily data from 1998 till present are too big to be downloaded to and processed on the users' premises, at least for the majority of the users. Note that early 2017 the full archives of SPOT-VEGETATION and Proba-V

were reprocessed to get one consistent time series. All reprocessed data and additional data (Level 2A) are available on the Proba-V MEP. On the platform different types of users have the tools for the tasks they intend to do, ranging from viewing the data till developing and operating user applications.

- On top of the EO-data mentioned above, the platform does co-locate as well complementary data in a way that it is easy accessible. Furthermore tools, libraries and applications which can be used by the large community are provided. The list of data and tools is constantly growing according user feedback and users can as well add these themselves.
- The platform can stimulate collaboration between the users, as we bring together services from various users on the same platform with a number of tools to support the publishing of and to provide feedback on these services. A further focus on documentation (blogs, tutorials, promotion of research done on the platform, etc.), knowledge sharing and user support complements this.
- As an Exploitation Platform (EP) with a focus on open interfaces, we position the Proba-V mission in an ecosystems of TEPs (Thematic EPs), REPs (Regional EPs) and other MEPs. We have the ambition to integrate the Proba-V MEP gradually in a federation of different platforms, as we do today in R&D context e.g. in the H2020 NextGEOSS, DataBio and TEP Food Security projects. In the future the DIAS platform, as intended to be operational by early 2018, offers a public IaaS (i.e. not the private cloud at VITO where the current Proba-V MEP is currently deployed) where the platform can be provided.

During the Proba-V MEP project, which will at least last till the end of the Proba-V mission in 2019, several third-party service projects are currently developing applications on the platform which can later on be offered as operational services on the same infrastructure. In the different platform development iterations, we address their user requirements and feedback to implement the shift of paradigm from “data to user” to “user to data”, bridging the gap between the traditional EO ground segment and the scientist or value added industry by providing a one stop shop for access to the full Proba-V Mission data (including derived parameters) and complementary data.

2. TECHNICAL SOLUTION

The Proba-V MEP provides scalable processing facilities with access to the complete data archive and a rich set of processing algorithms, models, open source processing libraries/toolboxes and public/collaborative software. The platform becomes the hub processing infrastructure of the mission by functioning as a powerhouse system and open access development environment.

To realise this the platform consists of the following components:

- The existing Product Distribution Facilities [4] and [5], are serving the access to the data archive, both via a Web portal as well as standardised discovery, viewing and data access interfaces. More evolutions on these standardised machine-to-machine interfaces are being developed e.g. the support of OpenSearch discovery and data access over HTTP with the possibility to customise products before downloading.
- Hadoop, as a software framework for data-intensive distributed applications, is designed to process large amounts of data by separating the data into smaller chunks and performing large numbers of small parallel operations on the data. Oozie is used as a workflow processing engine to design an EO-application as a workflow of multiple processes. Spark is used intensively on the MEP to allow analytics on large time series of data. The Hadoop ecosystem provides furthermore a rich and still growing set of tools which are used to give fast access to the data in a format needed by the specific application. As an example Accumulo and Geotrellis are used to offer data analytics on the large time series for user-defined polygons or single pixels as part of the Time Series Viewer.
- All EO raster data is accessible via NFS and possibly uploaded to the Hadoop Distributed Filesystem (HDFS) using a DataManager which integrates with several catalogues implementing different protocols, so that as well third party-data can be ingested in the platform when needed by a specific user.
- Cloud computing technology enables dynamic resource provisioning and is therefore providing a performing and scalable solution. OpenStack is chosen as private cloud middleware. Pre-configured virtual machines are offered and can run on the OpenStack cluster at VITO, providing the environment needed for users to work with the data and develop/deploy applications on the platform, i.e. containing IDE's, a rich set of tools and access to the complete data archive. Several external

users are currently performing R&D on these VM's, as explained in section 3.

- Interactive Web-based dashboards are designed to provide user-tailored information from the EO-data archives of VITO and other providers, by combining existing components such as AngularJS, Javascript libraries and GIS components into one single solution. As an example, the Proba-V MEP Time Series viewer allows you to view time series for any pixel of user-defined polygon for Proba-V data, derived vegetation indices and meteo data. Remark that the derived vegetation indices are originating from the Copernicus Global Land Service.

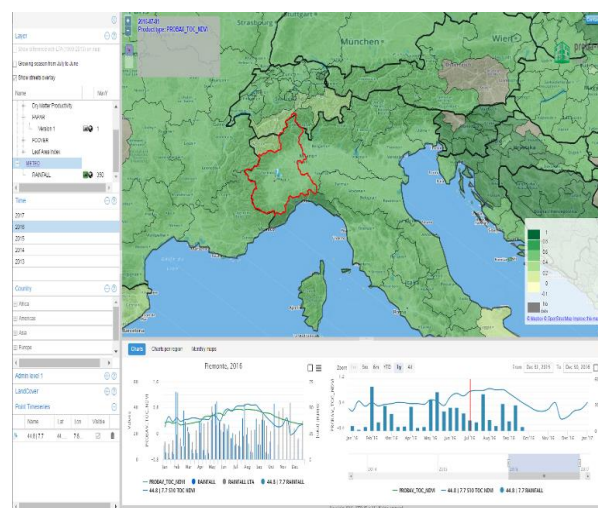
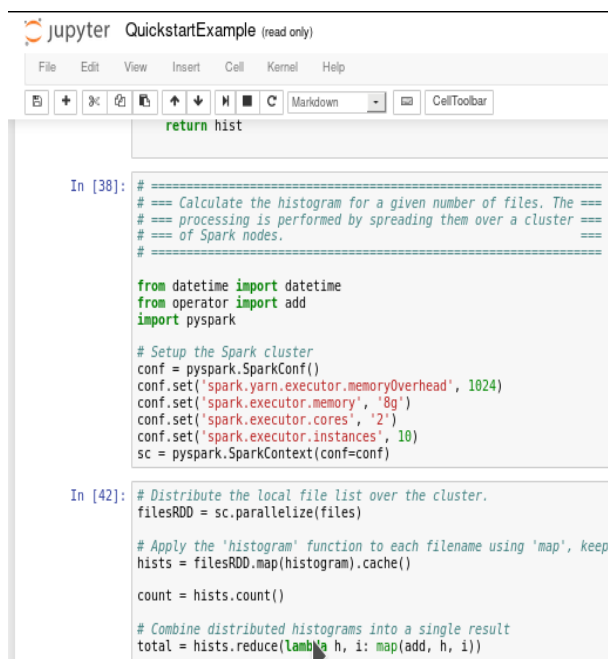


Figure 1: The Proba-V MEP Time Series viewer

- The Jupyter Notebooks Web application lets you create and share documents that contain live code, equations, visualisations and explanatory text. It is based on the Open Source Jupyter notebooks application, and tailored to the needs of remote sensing users. For programming, users can choose between the Python and R programming languages and can work interactively with the full data archive available at the Proba-V MEP. An ever growing list of software libraries such as GDAL, rasterio, pandas, numpy, matplotlib and seaborn is included and users can also upload and install their own packages and file. The Proba-V MEP Web portal provides several example notebooks, showing how to access data, how to use the time series viewer, how to plot charts and maps.



```

jupyter QuickstartExample (read only)
File Edit View Insert Cell Kernel Help
return hist

In [38]: # =====
# == Calculate the histogram for a given number of files. The ==
# == processing is performed by spreading them over a cluster ==
# == of Spark nodes. ==
# =====

from datetime import datetime
from operator import add
import pyspark

# Setup the Spark cluster
conf = pyspark.SparkConf()
conf.set('spark.yarn.executor.memoryOverhead', 1024)
conf.set('spark.executor.memory', '8g')
conf.set('spark.executor.cores', '2')
conf.set('spark.executor.instances', 10)
sc = pyspark.SparkContext(conf=conf)

In [42]: # Distribute the local file list over the cluster.
filesRDD = sc.parallelize(files)

# Apply the 'histogram' function to each filename using 'map', keep
hists = filesRDD.map(histogram).cache()

count = hists.count()

# Combine distributed histograms into a single result
total = hists.reduce(lambda h, i: map(add, h, i))

```

Figure 2: Python notebook example from Proba-V MEP

- A Web portal provides access to all applications and tools offered by the Proba-V MEP and to the cloud consoles. Furthermore the portal provides all information on the data and components available on the platform and offers tools for e-collaboration and knowledge sharing amongst the users. Blogs and tutorials are added regularly to respond to technical questions from users. In the next development iteration, starting in November 2017, a lot of new blog articles and sample projects will be added on the Web portal, together with user success stories which were performed on the Proba-V MEP.
- A main concern in the architecture was security since we allow user to develop and execute their applications on the platform. Their IPR shall be properly protected and the activities of individual users cannot influence the stability of the system and the work of other users. Single sign-on and proper monitoring of used resources are further requirements.

3. USERS AND FUTURE USE

Since the first pre-operational release, the Proba-V MEP is used by beta-testers to provide early feedback. Several users are developing a processing workflow or porting an application on the Proba-V MEP in the frame of the ESA MEP-TPS project. Often the Proba-V MEP is hosting the data intensive backend service, while the frontend remains at the premises of the user. The users range from universities to SMEs from different European countries, as listed in <https://proba-v-mep.esa.int/proba-v-mep-tps-users>.

Up to know, approximately 50 users did request a virtual machine or notebook access on the platform, including as well several VITO researchers, which indicates that the platform is as well realising the paradigm shift to bring the users to the data in the daily work of the Remote Sensing researchers and application developers of VITO, in the frame of several international projects.

Another user who performed a significant processing on the platform is the Copernicus Global Land Service lot-1, who created land cover maps for Africa processing all available Proba-V 100 m data. The metrics extraction needed +5000 CPU-hours for whole Africa and was processed on the Proba-V MEP with 517 executors in 22 hours. That's 99.6% time saving!



Figure 3: False color composite showing the biomass density, derived from Proba-V 100 m . The darker is more biomass. This was calculated on the Proba-V MEP in the frame of the Copernicus Global Land Service lot-1.

The usage of the Proba-V MEP applications is increasing constantly from the early release of the platform. The geoviewer, providing a full-resolution viewing service is as well intensively used in the promotion activities for the Proba-V user segment, e.g. in the weekly image-of-the-week. The time series viewer is used by several researchers, but as well by educational users. E.g. an Earth Observation course was developed by the Belgian office of the European Space Education Resource Office (ESERO) which uses intensively the Proba-V MEP time series viewer to illustrate long-term changes in vegetation e.g. deforestation or droughts.

The first on-demand processing service on the platform, the N-daily compositor application, is used by several Proba-V users to compute N-daily composites of Proba-V 1 km, 300 m or 100 m data with a sliding window for any area of interest. E.g. a user can request to get a 7-daily composite at 100 m spatial resolution every Monday over a given area of interest, and is no longer depending on the standard products offered by the Proba-V user segment. Furthermore different compositing algorithms are provided.



Figure 4: View the results from the on-demand processing service, before downloading

The Proba-V MEP is as well used in several R&D and pre-operational projects as a node within a federation, as explained earlier. E.g. in the EC H2020 NextGEOSS [8] and DataBio [9] projects, the Proba-V MEP is used to provide access to the full Proba-V archive and Sentinel-2 data over limited areas, towards researchers and on-demand pre-operational processing chains. In the ESA TEP Food Security project [10], the Proba-V MEP is serving the data analytics capabilities of the project, in a federation with public clouds such as IPT Poland.

Recently, VITO is assigned as Designated Entity to develop the Belgian Sentinel Collaborative Ground Segment (CGS), named Terrascope. The Proba-V MEP solution will hence be extended to adopt Sentinel-1/2/3 data and derived

products into a multi-mission platform, providing easy access to these data via WMS, WMTS and WCS interfaces and a cloud-based processing platform where users can upload-develop-test their own algorithms. Furthermore other atmospheric and geometric correction algorithms will be applied and 4 vegetation parameters NDVI, fAPAR, fCOVER and LAI will be provided. However, currently the scope of the Belgian CGS is being refined in a process of user consultation. As well the geographic coverage is still discussed, ranging from country scale for Sentinel-1, continental scale (Europe and Africa) for Sentinel-2 and global scale for Sentinel-3 to extend the time series from SPOT-VEGETATION and Proba-V.

3. CONCLUSION

The platform is fully operational since early November 2016 and is accessible from <https://proba-v-mep.esa.int/>. Two more development iterations are planned to further expand the capabilities of the system and provide new features, in close collaboration with the first third-party projects working on the platform. In 2017 several updates will be done, e.g. the support of Jupyter Notebooks and a better job management system for developers on our Hadoop platform.

The impact of this Proba-V MEP on the user community will be high and will completely change the way of working with the data and hence open the large time series to a larger community of users. During the presentation the capabilities of the platform will be demonstrated and an outlook for future developments and collaborations will be given.

4. REFERENCES

- [1] <http://proba-v.vgt.vito.be/>.
- [2] <http://land.copernicus.eu/global/>.
- [3] <https://proba-v-mep.esa.int>.
- [4] <http://www.vito-eodata.be>.
- [5] <http://land.copernicus.vgt.vito.be/PDF/>.
- [6] Proba-V Mission Exploitation Platform, Remote Sensing Journal, Technical Note, 2 July 2016, <http://www.mdpi.com/2072-4292/8/7/564/pdf>.
- [7] Proba-V MEP Leaflet for developers, https://proba-v-mep.esa.int/sites/proba-v-mep.esa.int/files/documents/mep_fact.sheets_finalweb.pdf.
- [8] <http://nextgeoss.eu/>.
- [9] <https://www.databio.eu/en/>.
- [10] <https://foodsecurity-tep.eo.esa.int/>.

PRESERVATION AND HARMONIZATION OF HISTORICAL AVHRR LAC DATA TO SERVE THE NEEDS OF USERS IN CLIMATE RESEARCH

Stefan Wunderle¹, Christoph Neuhaus², Fabia Hüsler¹, Andrew Brooks³, Neil Lonie³, Mirko Albani⁴, Sergio Folco⁴, Rosemarie Leone⁴

¹ Oeschger Center for Climate Change Research (OCCR), Department of Geography, University of Bern, Switzerland.

² Department of Geography, University of Bern, Switzerland

³ Dundee Satellite Receiving Station (DSRS), University of Dundee, UK

⁴ ESA-ESRIN, Frascati, Italy

ABSTRACT

Historical satellite data are of high value for climate research as an independent source to validate the output of climate models. Of special interest are data from the AVHRR sensor in 1km spatial resolution (LAC), which are available from 1981 until 2017 (2022). The data archived at different centers in Europe need consolidation (filling data gaps, removing redundant orbits etc.) and harmonization (generating the same format and meta-files) before transferred to ESA facilities to make a long time series accessible for scientific use. In the frame of ESA's Long Term Data Preservation program the archived AVHRR data will be stored and maintained for the next +50 years. We will present the different parts of the ESA-UniBern AVHRR-LTDP project and recommendations for the next processing steps needed to generate essential climate variables (e.g. snow extent, lake surface water temperature).

Index Terms— AVHRR-LAC data, ESA Long Term Data Preservation (LTDP), data consolidation and harmonization

1. INTRODUCTION

During the development phase of the AVHRR sensor in the early 70ies, no one could imagine that the system would become very attractive for climate research. This is due to its almost unchanged sensor during the last 35 years on the NOAA-series that makes it the longest time-series of optical satellite imagery in a daily resolution. Furthermore, with the cooperation between NOAA and EUMETSAT, under the Joint Polar Satellite Systems (JPSS) agreement, the sensor will be in orbit until 2022 on MetOp. A unique data set for climate research covering a period from 1981 – 2022, which is the basis for many investigations, in addition to being used as a complementary and independent source to validate climate models. While many projects use the globally available AVHRR Global Area Coverage (GAC) data with a degraded spatial resolution, higher spatial resolution is mostly needed. Therefore, the AVHRR Local Area Coverage

(LAC) data with a ground sampling distance (GSD) of 1.1km in nadir are of special interest. Due to limited on-board storage capability of the NOAA-satellites the availability of LAC data relies on local receiving stations and archiving facilities. Depending on financial resources and used hardware and software the archived data exhibit temporal and spatial data gaps, different file formats (HRPT, NOAA-level 1a, SHARP, etc.) and processing levels, different archiving and storage techniques. In addition, many of the local archives at Universities are managed as a “one-man-show” relying on enduring interest of a few scientists. Contrarily, professional data centers or meteorological services had the resources to keep the data useable but political decisions often resulted in a complete loss of AVHRR data, especially from the years 1981 – 1995. Hence, there is a strong need by scientists to get access to a homogenized and consolidated AVHRR LAC archive.

2. DATA AVAILABILITY

Under the circumstances of the financial and administrative restriction mentioned above it is remarkable that University of Bern (UoB) has one of the longest AVHRR LAC archives (1989 – 2017), which is accessible for scientific use without any restrictions (figure 1). Covering whole Europe in a daily resolution the UoB archive was defined as a European Master Data Set.

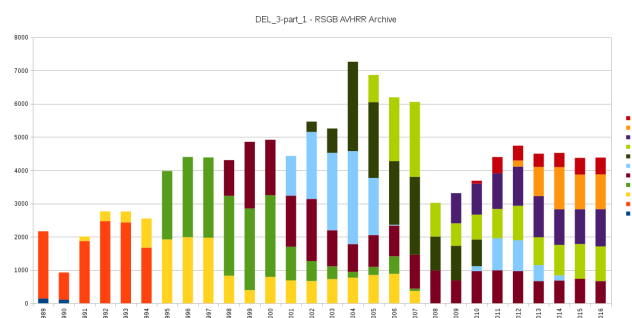


Figure 1: archived AVHRR LAC data at UniBern

In the framework of ESA's Long Term Data Preservation (LTDP) program the AVHRR LAC data received and archived at ESA facilities (Maspalomas, Frascati, Tromso) are read from old tapes and stored in fast accessible archives (figure 2).

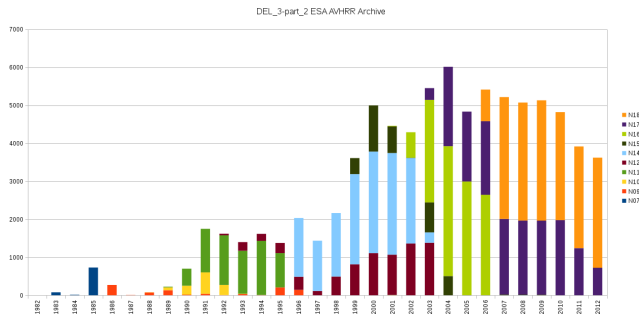


Figure 2: archived AVHRR LAC data at ESA facilities

The aim of ESA-UoB project AVHRR-LTDP is to use the UoB European Master Data Set and to combine it with ESA-AVHRR data. Furthermore, Dundee Satellite Receiving Station (DSRS) has systematically received and archived AVHRR LAC data since the first sensor was launched in the late 1970's. It maintains extensive and continuous archives (1978 - 2017) which will be used to extend and supplement the European Master Data Set.

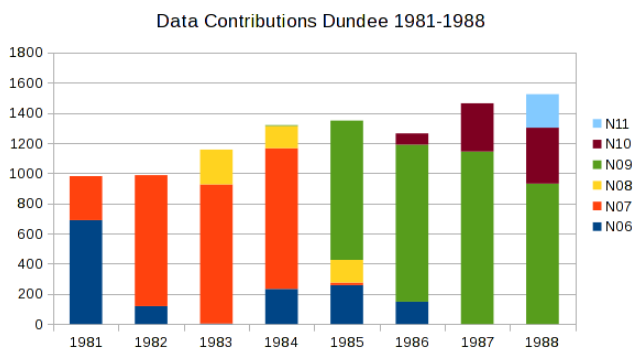


Figure 3: Subset of AVHRR LAC data archived at DSRS (1981 – 1988) to be included in the harmonized AVHRR data archive.

After consolidation and harmonization of the time series all of the data will be transferred to ESA facilities to make them accessible for the next 50+ years.

3. PRESERVATION OF AVHRR DATA

It has been decided to compile a data archive based on data of the lowest level (raw data) to maintain the option of extracting all necessary meta data useable for quality checks etc. Hence, the first step of the LTDP-AVHRR project was an analysis of all available data sets at ESA facilities and UoB

in terms of satellite, coverage, temporal resolution and data format. This information formed the basis to detect gaps at the UoB Master data set and to screen the ESA holdings for potential gap filling. Only a few minor data gaps for the period 1989 – 2016 were found. However, the different formats at UoB (HRPT, NOAA level 1b) and ESA (SHARP) may cause some additional effort in data processing / re-formatting. Moreover, there is a need to fill the previous years, at least from 1985 – 1989, to fulfill the WMO requirement of a climate period (30 years minimum), which offers climate modelers a long time series for sound statistical analysis. The only archive with the needed AVHRR LAC data of this period is that of the Dundee Satellite Receiving Station (DSRS) – agreement has been reached for DSRS to support the project and include also its data in the European Master Data Set. In the framework of the project the content of the archives is documented to start re-processing for harmonization.

4. HARMONIZATION

Some of the re-processing steps were defined with support of an established AVHRR advisory group. This included decisions on the final data format, stitching of images from same orbit but with different length, exclude identical data sets in the user accessible area but keep the image as backup, homogenize file names, etc. After these decisions the reprocessing at UoB Processing and Archiving Facility (PAF) has started. The needed ESA data are transferred to UoB to guarantee a consistent reprocessing of all data (> 200.000 data sets/scenes) and generate quality flags, Quicklooks and all meta-data needed to compile the final archiving format at ESA (EO-SIP).

5. STEPS TOWARDS SERVING THE NEEDS OF SCIENTISTS IN CLIMATE RESEARCH

In previous studies related to the usability of AVHRR data many users pointed out their interest in long time series of products (essential climate variables) and their limited capability to process raw AVHRR data. Hence, a careful pre-processing including calibration and geocoding is foreseen to be the next step after the compilation of the European AVHRR Master data set.

5.1. Calibration

The AVHRR calibration process of the shortwave and thermal channels varies with respect to the availability of adequate calibration information [1]. For the thermal channels, an onboard calibration information based on a view of stable blackbody and deep-space reference is provided, which is utilized to convert raw counts to a meaningful physical quantity: the brightness temperature. In contrast to the thermal channels, the visible and near-infrared channels are only calibrated pre-launch, which complicates the calibration procedure as their signal was observed to decrease

over time. To account for this fact, time-dependent correction using updated calibration coefficients is required. Here, we currently use the updated coefficients suggested by Heidinger et al. 2010 [2] and Heidinger et al. 2014 [3]. Based on these updated coefficients, channels 1, 2 and 3A are calibrated to top-of-atmosphere (TOA) reflectance as described in the NOAA User's Guides.

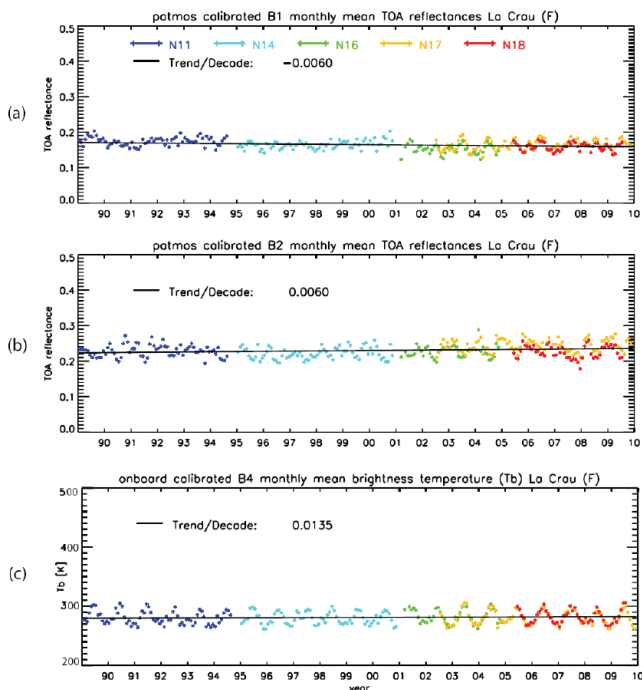


Figure 4: Time series of mean monthly TOA reflectance (a (channel 1), b (channel 2)) and brightness temperature of channel 4 (c) of La Crau. (Hüsler et al. 2011)

5.2. Geocoding

Accurate geolocation remains critical for generating long-term data records for climate studies of land surface and atmospheric parameters. Inaccurate co-registration of consecutive images may lead to biases in time series since every single scene contributes to the final product.

The implemented geocoding and orthorectification procedure SAPS (Science Systems and Applications, Inc. AVHRR Processing System) at University of Bern has been developed by Khlopenkov et al. 2010 [5]. The processing system relies on 250m MODIS monthly composites as reference images and 500m grid spacing SRTM digital surface model. The final accuracy of the AVHRR images after geocoding/orthorectification is 1/3 pixels on average [1].

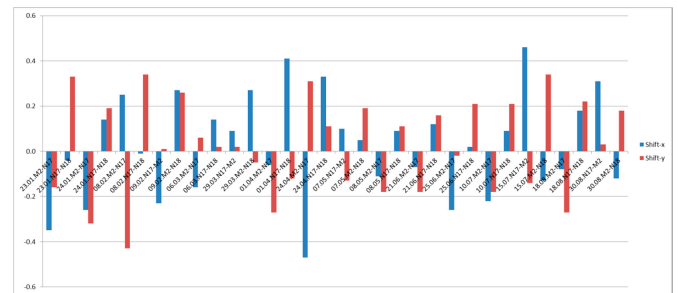


Figure 5: Mean x-y-shift (pixel) from 32 image pairs (NOAA-18, NOAA-19 and MetOp-A). from Aksakal et al. 2015

5.3. Product retrieval

The benefit of long time series based on AVHRR data will be shown at the end of the presentation with two examples: Snow Extent (SE) for the European Alps and Lake Surface Water Temperature (LSWT) of some smaller lakes in Europe.

6. CONCLUSION

The combination of three major AVHRR data archives in Europe, namely University of Bern, European Space Agency and Dundee Satellite Receiving System will result in a homogeneous and consolidated time series with high impact in climate change studies. The final data set covers Europe from 1980 – 2017 with daily observations and a spatial resolution of 1km (nadir). Some effort is needed to generate a useful level 1c data set including calibration and geocoding.

7. REFERENCES

- [1] Aksakal, Sultan; Neuhaus, Christoph; Baltasvius, Emmanuel; Schindler, Konrad (2015). Geometric Quality Analysis of AVHRR Orthoimages. *Remote sensing*, 7(3), pp. 3293-3319. Molecular Diversity Preservation International MDPI 10.3390/rs70303293
- [2] Hüsler, Fabia; Fontana, Fabio; Neuhaus, Christoph; Riffler, Michael; Musial, Jan; Wunderle, Stefan (2011). AVHRR Archive and Processing Facility at the University of Bern: A comprehensive 1-km satellite data set for climate change studies. *EARSeL eProceedings*, 10(2), pp. 83-101. Oldenburg: BIS Verlag
- [3] Heidinger, A.K., William C. Straka III, Christine C. Molling, Jerry T. Sullivan & Xiangqian Wu (2010): Deriving an inter-sensor consistent calibration for the AVHRR solar reflectance data record, *International Journal of Remote Sensing*, 31:24, 6493-6517;
- [4] Heidinger, A. K., Foster, M. J., Walther, A., and Zhao, X.: 2014. The pathfinder atmospheres-extended AVHRR climate dataset, *Bulletin of the American Meteorological Society*, 95, 909-922.
- [5] Khlopenkov, K.V., A.P. Trishchenko, Y. Luo (2010): Achieving subpixel georeferencing accuracy in the Canadian AVHRR processing system. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (4) (2010), pp. 2150-2161.

EXPLOITATION OF ENVISAT ASAR AND SENTINEL-1 SAR DATA IN SUPPORT OF CARBON AND WATER CYCLE STUDIES

Maurizio Santoro¹, Oliver Cartus¹, Andreas Wiesmann¹, Urs Wegmüller¹, Josef Kelldorfer², Christiane Schmullius³, Pierre Defourny⁴, Olivier Arino⁵, Marcus Engdahl⁵, Frank Martin Seifert⁵

¹ GAMMA Remote Sensing AG, Gümligen, Switzerland

² Earth Big Data LLC, Woods Hole, USA

³ Friedrich-Schiller University, Department of Earth Observation, Jena, Germany

⁴ Université Catholique de Louvain, Earth and Life Science Institute, Louvain-la-Neuve, Belgium

⁵ ESA ESRIN, Frascati, Italy

ABSTRACT

Spaceborne SAR dataset have been available for almost 30 years. C-band data records are the longest available; in particular data records from Envisat ASAR and Sentinel-1 are publically available, thus fostering the development of large-scale data products. In this paper, we review the processing behind the generation of SAR data stacks and related data products in support of applications in the domain of the carbon cycle and the water cycle. C-band SAR data were found to advance spatial detail and accuracy of estimates with respect to existing datasets; nonetheless, there are clear limitations that can be overcome with the synergy with other sources of observations from space.

Index Terms— SAR, Envisat ASAR, Sentinel-1, data preservation, data products.

1. INTRODUCTION

Synthetic Aperture Radar (SAR) data have been collected in regular manner during the last three decades. The independence of image acquisition from cloud cover and solar illumination allows accurate planning of frequency of observations in order to support mapping based on multi-temporal observations and monitoring to track dynamics. C-band satellites operated by ESA are the spaceborne SAR missions with the longest time record of observations (since 1991 with interruptions in 2001 and 2013). In addition, the data since 2002 are publically available, favoring the development of innovative mapping techniques and the continuation of existing mapping endeavors.

During the Envisat mission, the Advanced SAR (ASAR) was operated in multiple modes; in particular the ScanSAR mode provided repeated observations at moderate and coarse resolution (150 m and 1,000 m). For example, almost daily observations were collected over polar and boreal regions. The redundancy of observations has been exploited in studies supporting the assessment of variables of the

carbon [1] and the water cycle [2, 3] at regional scale. The importance of repeated observations has been implemented in the observation strategy of the Sentinel-1 mission. With respect to the Envisat mission, Sentinel-1 observations are more targeted to exploit as much as possible the information contained in the SAR data with respect to the thematic applications deemed feasible with C-band.

In this paper, we report on achievements with the records of Envisat ASAR and Sentinel-1 observations in support of the carbon and water cycle. Such achievements are based on the exploitation of multi-year observations of the SAR backscattered intensity available through data archives at the European Space Agency and the Alaska Satellite Facility.

2. SAR DATA PROCESSING ON GRID AND CLOUD COMPUTING FACILITIES

Classification and retrieval algorithms based on SAR data require that these are calibrated, co-registered and filtered for speckle noise. As data provided by space agencies or data archives are not available in such format, a processing scheme starting from the original data is required. In addition, such processing scheme needs to be performing on large datasets (software and hardware requirement). With the GAMMA Software [4], it was possible to easily implement the processing scheme to achieve the required data products ensuring an overall fast processing turnaround with high quality data products certified by flags and processing statistics.

Back in 2010, the Grid Processing on Demand (G-POD) facility at ESA was the first of its kind to provide access to long-term archives of SAR images and a processing facility on which a processing sequence could be easily implemented. G-POD was set up to be a generic GRID-based operational environment allowing access to processing facilities closer to Earth Observation (EO) data stored locally. For this, software code could be implemented on the platform to easily access the data and run the code.

On G-POD, Envisat ASAR images acquired at moderate resolution (i.e., 75 m and 150 m) over all land masses except Antarctica between 2005 and 2012 were processed (more than 200,000 images) to form stacks of co-registered images of the SAR backscatter. The objective was to have at hand time series of observations to be used in multiple thematic applications. In a second exercise, all ASAR images acquired between October 2009 and February 2011 (30,000 images) were processed to 1,000 m pixel size for the scope of generating a map of forest biomass of the northern hemisphere representative for the year 2010.

For the processing, a Computing Element consisting of 10 processing nodes and based in ESRIN was made available. The GAMMA software was installed on the platform and processing chains consisting of commands of the software arranged in batch scripts were implemented on G-POD. A user interface allowed selection of data to be processed and managing the processing queue. Pre-processing consisted of data ingestion, calibration, terrain geocoding to a geographic coordinate system, tiling in windows of $1^\circ \times 1^\circ$ (for 150 m data) or $2^\circ \times 2^\circ$ (for 1,000 m) and multi-channel speckle filter. Because of limited resources available at that time, a processing task could at most handle 100 ASAR moderate resolution images or 1,000 ASAR coarse resolution images. Processing of the global ASAR dataset at moderate resolution resulted in over 2,000 tasks. Data selection for each tasks was extremely cumbersome because of redundancies in the ASAR archives at ESA, requiring manual de-selection of copies of an original image dataset. The pre-processed data consisted of almost 40 Terabytes of data including images of the SAR backscatter, the local incidence angle and the pixel area. Data were stored locally on external hard disks and shipped to GAMMA once full.

The G-POD exercise served to lift such a system to cope with the demands of big data processing, teaching a number of lessons to both service providers and users.

With the increased availability of grid and cloud computing facilities responding to the need of the larger throughput of data, e.g., by spaceborne remote sensing platforms, we utilized Earth Big Data's Cloud Processing system to process time series data of Sentinel-1 images covering West Africa, acquired between October 2014 and May 2017 (approximately 6,300 images). Earth Big Data LLC (EBD) has developed a fully automated processing solution that ties SAR processing software (e.g., GAMMA) and geospatial open source processing tools (like GDAL, python, R, Octave, etc.) into a scalable cloud processing environment utilizing Amazon Web Services (AWS).

Backbone of the processing is the EBD Software for Earth Big Data Processing, Prediction Modeling and Organization (SEPPO). At its core, SEPPO features a flexible "recipe processor" that can ingest a range of data from various cloud resource endpoints at run time and can be coupled with the appropriate machine size of the AWS

elastic compute instance fleet. Recipes are stand-alone processing algorithms that assume an input data stack to be available in a working directory in which the recipe performs all computations on the data. At conclusion of a recipe run, simple rules for transfer of the desired end products to permanent or cached cloud storage is performed. Data input and transfer to the final storage is implemented to work with standardized URL prefix strings that determine where input and output data are located including transfers via `scp` or `ftp` protocols. For example the "`s3://`" prefix determines to retrieve from or put data to the AWS simple storage system `s3` with AWS tools. The "`http(s)://`" prefix determines endpoints using the transfer protocols and software, including credential verification to obtain data from the respective URL location.

For the processing of the 6,300 raw Sentinel-1 GRD frames for this task, we designed a suite of recipes that would first perform radiometric terrain correction (RTC) based on the GAMMA software. This recipe obtained data from the Alaska Satellite Facility (ASF) obtaining the appropriate '`http://`' URL's from a spatial query of EBD's cloud-based Spatial Metadata Database (SMDDDB). The SMDDDB is updated nightly with holdings of data in the ASF archive and allows flexible query and intersections with region of interest data sets. Each of the 6,300 scenes was thus coupled with the appropriate processing recipe and required machine configuration, and submitted to the SEPPO implemented cloud processing queuing system. After queuing, 30 compute instances were launched in parallel to consume the processing queue. We limited the number of compute instances to 30 after consultation with the ASF to optimize data access rates. Recent trials suggest that parallel access with as many as 100 processors become quite feasible. As ASF is in the process of moving data holdings to the AWS `s3` store in the US-East-1 availability zone, we were able to achieve for the bulk of the data transfer rates of up to 80MB/s, which allowed us to access full Sentinel-1 GRD scenes of about 1GB in size in less than 20 seconds for a scene. RTC processing was queued to progress from West to East. All RTC products were stored in a AWS `s3` scratch space.

A feature of the RTC processing recipe was the generation of quicklook data at reduced resolution. Throughout the RTC processing, quicklook mosaics of entire Sentinel-1 strips were generated automatically and visually inspected for potential problems (Figure 1). Except for a handful of inaccessible GRD products at ASF, virtually no issues with the RTC processing was found. In parallel to the runs for RTC processing, we deployed a developed recipe for Sentinel-1 time series processing that included subsetting of Sentinel-1 scenes into $1^\circ \times 1^\circ$ geographic tiles of 0.000181818 arcsecond resolution (ca. 20 meters) covering the target region. This recipe performed multi-temporal speckle filtering on the subsetted data, incidence angle map generation for each distinct Sentinel-1 path

covering the tile, and output of the filtered time series and incidence angle stack to a zipped s3 archive file.

A total of 328 tiles covered the region of interest (Figure 2). For tile processing, larger AWS compute instances were required and requested via SEPO's queuing system. The consumption of the tile processing queue was coordinated via SEPO's implemented process completion dependency checks with the processing of the RTC products. As such, once all RTC products pertaining to the coverage of a $1^\circ \times 1^\circ$ tile (as determined from SMDDDB spatial intersection queries) were completed, tile processing of a respective tile kicked in.

To match the progress of RTC processing with tile processing, we deployed 10 tile recipe processors in parallel, each storing results in a zip data archive on AWS s3. After completion of the 328 tiles, dedicated data access URLs were generated and securely shared with the collaborators at Gamma Remote Sensing for data download. Total production time from RTC, processing, quicklook inspection, and tile processing was about 88 hours. Total input raw data volume of GRD scenes was about 6.2 Terabytes. Volume of final tile products was about 2.5 Terabytes.

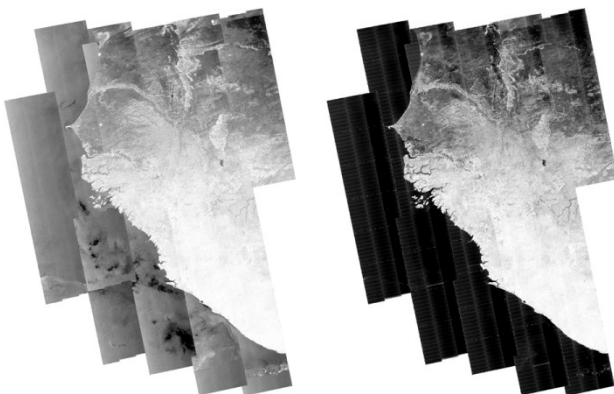


Figure 1. Strips of Sentinel-1 RTC Quicklook data for quick visual inspection of data quality. Left: VV polarization; Right: VH polarization.

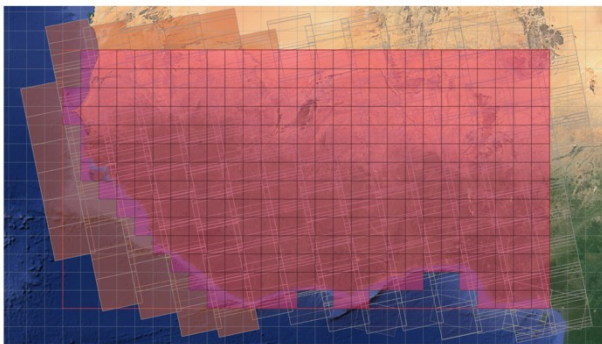


Figure 2. Tiles (red) in the target region and Sentinel-1 frame coverage (gray outlines). Orange colors show the progress of processing from east to west.

3. FOREST BIOMASS RETRIEVAL

The C-band SAR backscatter is not the most suited observable to retrieve forest biomass given its weak sensitivity in high biomass. Nonetheless, it was shown that combining hundreds of observations it is possible to improve the retrieval accuracy beyond what achieved with a single observation [1]. The BIOMASAR retrieval algorithm was therefore developed to implement such approach so that ultimately biomass is estimated from a set of hyper-temporal C-band SAR observations [1]. The algorithm was applied to the Envisat ASAR dataset acquired on a 17-months' time period about the year 2010 and processed on G-POD to retrieve forest biomass of the northern hemisphere [5]. This data product (Figure 3) is unique in terms of coverage, spatial detail (1,000 m spatial resolution) and thematic consistency, allowing assessments of biophysical processes of boreal and temperate forests [6]. The retrieval has recently being extended to tropical and sub-tropical forests to support the generation of a global dataset of forest biomass for the year 2010.

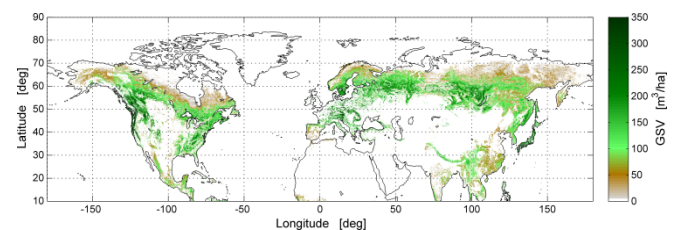


Figure 3. Estimates of forest biomass expressed as growing stock volume (GSV; unit m^3/ha) obtained from a 17 months dataset of Envisat ASAR backscatter observations around the year 2010 (reproduced from [5]).

4. MAPPING AND MONITORING OF INLAND WATER BODIES

Detection of water bodies in C-band SAR images is rather straightforward because of the very low backscattered signal. Nonetheless, this feature is not unique to water surfaces as sand dunes, very arid terrain and wet snow present the same signature. To reduce the water commission error, additional features need to be exploited. The time series of observations processed for the period 2005-2012 (only high- and moderate-resolution modes) allowed the generation of multi-temporal metrics with better separability between water and other land surface types. Ultimately, such SAR dataset was used to generate an indicator of water bodies [7] by means of a simple thresholding algorithm. The water bodies indicator with a spatial resolution of 150 m was selected as major dataset to support the generation of a global dataset of inland water bodies representative for the year 2010 [8]. The Climate Change Initiative (CCI) Land Cover (LC) water bodies product is a self-standing dataset of the CCI-LC climate data package (Figure 4).

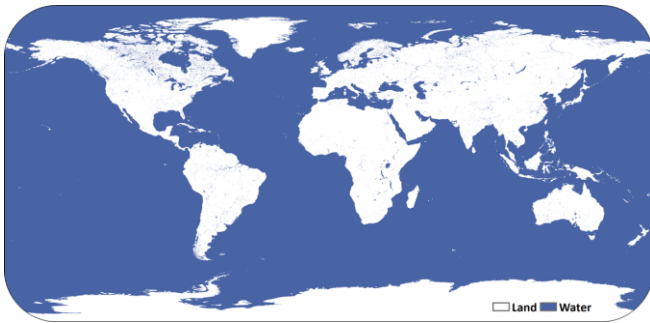


Figure 4. CCI Water Body dataset available at <http://maps.elie.ucl.ac.be/CCI/viewer> [8].

The entire time series of ASAR observations (from high- to low-resolution) has recently been used to generate an additional data product related to the water cycle. A climatology of water occurrence for 2005-2012 was generated with a spatial resolution of 1,000 m (Figure 5). The dataset portrays variability of water cover as seen by the ASAR instrument and is currently being assessed in terms of its thematic accuracy.



Figure 5. Occurrence of open standing water for the last week of May for the time period 2005-2012 in North America.

The data record of Sentinel-1 observations over West Africa has a unique feature of one observation every 12 days ensuring the possibility of building on the experience of mapping water bodies with Envisat ASAR data. Again, multi-temporal metrics have been exploited but a more sophisticated approach has been selected (bagged decision trees) to cope with the variability of the backscattered signal throughout the year. The Sentinel-1 water bodies dataset has a spatial resolution of 20 m and provides a complete representation of open water bodies of West Africa, not yet achieved with optical imagery (Figure 6).

5. ACKNOWLEDGMENTS

The work here presented was undertaken within the BIOMASAR (ESRIN contract No. 21892/08/I-EC), the GlobBiomass (ESRIN contract No. 4000113100/14/I-NB) and the Climate Change Initiative Land Cover (CCI-LC)

(ESRIN contract No. 4000101774/10/I-LG) projects, all sponsored by the European Space Agency (ESA). Access to the G-POD processing facilities was possible through ESA's Category 1 Project ID 9209 "G-POD processing of ASAR Wide Swath imagery for multi-purpose applications".

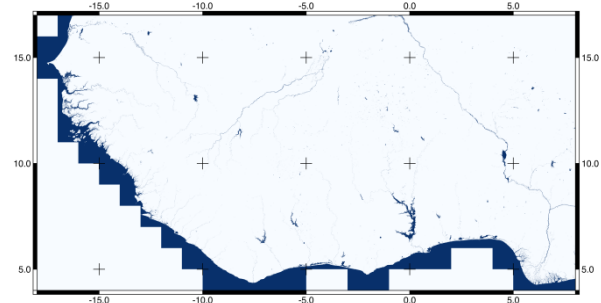


Figure 6. Map of water bodies of West Africa generated from time series of Sentinel-1 SAR backscatter observations.

6. REFERENCES

- [1] M. Santoro, C. Beer, O. Cartus, C. Schmullius, A. Shvidenko, I. McCallum, U. Wegmüller, and A. Wiesmann, "Retrieval of growing stock volume in boreal forest using hyper-temporal series of Envisat ASAR ScanSAR backscatter measurements," *Remote Sens. Environ.*, vol. 115, pp. 490-507, 2011.
- [2] A. Bartsch, R. Kidd, C. Pathe, W. Wagner, and K. Scipal, "Satellite radar imagery for monitoring inland wetlands in boreal and sub-arctic environments," *Journal of Aquatic Conservation: Marine and Freshwater Ecosystems*, vol. 17, pp. 305-317, 2007.
- [3] D. O'Grady, M. Leblanc, and A. Bass, "The use of radar satellite data from multiple incidence angles improves surface water mapping," *Remote Sens. Environ.*, vol. 140, pp. 652-664, 2014.
- [4] C. Werner, U. Wegmüller, T. Strozzi, and A. Wiesmann, "GAMMA SAR and interferometric processing software," *Proc. ERS-Envisat Symposium*, Gothenburg, 16-20 October, 2000.
- [5] M. Santoro, A. Beaudoin, C. Beer, O. Cartus, J. E. S. Fransson, R. J. Hall, C. Pathe, D. Schepaschenko, C. Schmullius, A. Shvidenko, M. Thurner, and U. Wegmüller, "Forest growing stock volume of the northern hemisphere: spatially explicit estimates for 2010 derived from Envisat ASAR data," *Remote Sens. Environ.*, vol. 168, pp. 316-334, 2015.
- [6] M. Thurner, C. Beer, M. Santoro, N. Carvalhais, T. Wutzler, D. Schepaschenko, A. Shvidenko, E. Kompter, B. Ahrens, S. R. Levick, and C. Schmullius, "Carbon stock and density of northern boreal and temperate forests," *Global Ecol. Biogeogr.*, vol. 23, pp. 297-310, 2014.
- [7] M. Santoro, U. Wegmüller, C. Lamarche, S. Bontemps, P. Defourny, and O. Arino, "Strengths and weaknesses of multi-year Envisat ASAR backscatter measurements to map permanent open water bodies at global scale," *Remote Sens. Environ.*, vol. 171, pp. 185-201, 2015.
- [8] C. Lamarche, M. Santoro, S. Bontemps, R. d'Andrimont, J. Radoux, L. Giustarini, C. Brockmann, J. Wevers, P. Defourny, and O. Arino, "Compilation and validation of SAR and optical data products for a complete and global map of inland/ocean water tailored to the climate modeling community". *Remote Sens.*, vol. 9, 36, 2017.

SYSTEM FOR AUTOMATIZED SENTINEL-1 INTERFEROMETRIC MONITORING

Milan Lazecký¹

¹IT4Innovations, VSB-TUO, Ostrava, Czechia

ABSTRACT

A preparation of database of pre-processed images is a prerequisite for effective analyzes using SAR interferometry (InSAR), but can be used also for intensity and polarimetry analyzes. Several implementations of the Permanent Scatterers (PS) InSAR technique demonstrated that interferograms do not necessarily have to be generated prior to the processing, especially if only temporal unwrapping is to be performed in the time series. Czech nation-wide database contains Sentinel-1 bursts that have been preprocessed to the state of a consistent well-coregistered dataset. The further processing time is significantly reduced in order to achieve PS or other, e.g. Small Baseline (SB) - based velocity maps. Every new pre-processed burst can also trigger a processing update that is able to detect unexpected changes from InSAR time series and therefore provide a signal for early warning against suspicious occurrence of a potential dangerous displacements. Work towards such early warning system is still ongoing, while the system running at IT4Innovations high performance computing facility (HPC) is already able to provide fast InSAR results over Czech territories.

Index Terms— Sentinel-1, SAR Interferometry, SAR Processing, HPC

1. INTRODUCTION

Copernicus Sentinel-1 SAR satellite system offers radar imagery of European areas every 12 days since autumn 2014 and every 6 days since autumn 2016. Its technical characteristics are very satisfying for InSAR applications. Own past works based on current InSAR research knowledge have proved the efficiency of Sentinel-1 InSAR analysis for identification of mainly vertical displacements, ranging from few millimeters per year such as displacements of bridges in Ostrava and Prague [1] or Plover Cove dam in Hong-Kong [2] to the range of centimeters per year in subsiding Konya city in Turkey [3] or decimeters per month in case of areal subsidence troughs in Karvina region [4]. In proper conditions, a motion of slope can be also identified making the techniques and the satellite itself useful for distinguishing between active and non-active landslides [5]. These works and findings were used as a proof of a unique and practical applicability of Sentinel-1 InSAR analyzes and were the base for a support

to establish a nation-wide Sentinel-1 InSAR monitoring system (IT4S1) at Czech national supercomputing center, IT4Innovations.

There are several methods of InSAR developed since the publication of Permanent Scatterers (PS) technique in 2000 [6]. Current PS implementations applied on Sentinel-1 data allow identification of near-vertical displacements in the rate of up to few decimeters per year with the standard deviation often around 1 mm/year. However the technique is applicable only on “clean” points, i.e. at least without the presence of vegetation in the observed location. For monitoring of natural areas, specific techniques were developed such as Small Baseline InSAR (SB) [7] or partially coherent PS InSAR [8]. All of these techniques start with focused Single Look Complex (SLC) SAR data with a radar phase component included, that are combined interferometrically. Situation with Sentinel-1 images is more complex due to usage of so-called TOPS mode. A significant disadvantage of TOPS mode is the necessity of Enhanced Spectral Diversity (ESD) correction [9] achievable only using large-scale portions of the Sentinel-1 images by combinations of overlapping sub-images taken from slightly varying observing angle called bursts. The IT4S1 system prepares separate bursts after coregistration, ESD correction and other phase (e.g. Elevation Antenna Patterns, EAP, where needed) or radiometric corrections as corrected SLC images (SLC-C) into a database offering a fast and effective post-processing, especially InSAR analyzes using some of mentioned techniques.

2. OBJECTIVES AND EXPECTED IMPACT

By establishing a system that is able of fast on-demand InSAR processing of current Sentinel-1 data, the society can achieve several products (see Figure 1):

- static annual maps of active slope failures (especially creeps or slow landslides) – this information is crucial for risk management actions. Risk management is often not aware about the threat of landslides in flooded areas where floods may activate the existing slope failure. Only sparsely updated information is available to risk managers from Czech Geologic Survey based on expert evaluation of current state of slopes. The InSAR-based maps can give an additional though experimental information raising the caution of landslide activity in affected areas.
- static maps of (vertical) displacements of structures, with a millimeter sensitivity – remotely acquired information about

current displacements can play an important role for identification of potential structure issues. As typical structures, one may mention transportation structures (roads, railroads, bridges), dam constructions, inhabited buildings, electricity towers etc. To achieve the most complete information, data from opposite satellite passes can be combined into an analysis known as a decomposition of line-of-sight (LOS) values into horizontal and vertical directions (further as decomposition).

- static (semi-)annual maps of terrain development in urban or non-urban areas, such as development of e.g. mine-induced subsidence. Provided information about identified terrain changes or a stabilization of movements in affected areas can be important information for e.g. municipal urban planning facilities.

- based on the burst SLC-C database, an early warning system can be arranged in a relatively straightforward way that would continuously update displacement values over critical infrastructure and raise attention to end-users for a verification (ongoing work).

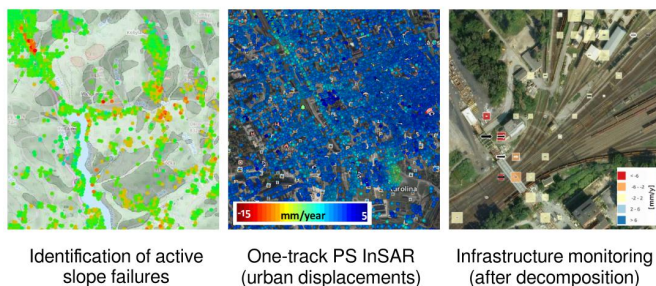


Figure 1. Application examples for on-demand processing of Sentinel-1 data based on selected area of interest.

3. SYSTEM ARCHITECTURE

An HPC-based system for a continuous pre-processing of all Sentinel-1 SLC images over Czechia is arranged, based on a connection to the Czech Copernicus Collaborative Segment (CollGS) maintained by CESNET organization. A new SLC image arrives to CollGS in less than 30 hours after image acquisition, however currently only data containing precise orbit ephemerides (POD) are updated in IT4S1 system that leads to 21 days update delay of new images.

The Sentinel-1 burst SLC-C database system consists of four hardware segments, see Figure 2. After arrival of new Sentinel-1 SLC image to CollGS, a LiCS solution [10] based metadata database system (a metadata base) ensures a proper identification of its bursts including information about their geographic location. Based on availability of POD, the system allows the given image to be partitioned to bursts, radiometrically calibrated by an SLC preprocessor server that sends the output to the SLC-C preprocessing HPC facility. Custom solution prepares coherent burst combinations in order to perform ESD computation. At this stage, ISCE algorithms (NASA JPL/Caltech) [11] are

applied, performing the burst preprocessing until the stage of generating range fine offset fields for every burst (containing phase estimation of several non-displacement phase signature sources). These offset fields are removed from the bursts, inducing InSAR-ready burst SLC-C images. This set of operations is the most computationally demanding.

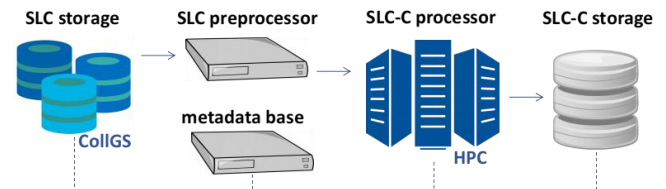


Figure 2. Architecture of IT4S1 system Sentinel-1 burst SLC-C generation.

The result is a database of bursts that are coregistered to each other in the precision of around 0.001 pixel that is necessary for a proper InSAR processing. The bursts SLC-C files can be easily combined on-demand by some or all of the implemented multitemporal InSAR techniques (currently PS and SB implementations based on STAMPS [12], the works towards inclusion of other implementations, such as SARPROZ are ongoing). Results can be post-processed and visualized in a webGIS interface (under preparation), showing to end-user full resolution mean velocity maps with a possibility of achieving time series figures for a selected point. The diagram of the multitemporal processing part is shown at Figure 3. The system is primarily designed for an on-demand InSAR analysis in order to identify ground or structure displacements. Since the SLC-C files are generated and stored, the computational burden in this part is minimized.

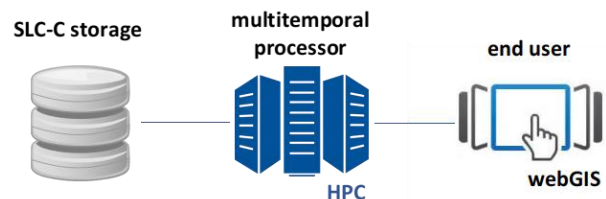


Figure 3. Architecture of IT4S1 system multitemporal InSAR processing and visualization.

4. COMPUTATIONAL RESOURCES

The IT4Innovations facility currently offers a supercomputing cluster Salomon. Salomon cluster consists of 1008 computational nodes of which 576 are regular compute nodes and 432 accelerated nodes. Each node is a powerful x86-64 computer, equipped with 24 cores (two twelve-core Intel Xeon processors) and 128 GB RAM. The nodes are interlinked by high speed InfiniBand and Ethernet networks. A DDN Lustre shared storage offers a capacity of 1.69 PB.

Sentinel-1 data cover the whole Czechia from 8 tracks, yielding ~160 new images per month. One image contains 24 bursts (8 bursts per 3 swath units) covering ~80x18 km area each with its LOS resolution of ~3x13 m (~5x20 m in ground range) and has data size of ~4.5 GB in its compressed form in distribution, i.e. ~200 MB/burst in the compressed form and ~550 MB in uncompressed form (including both co-polarized and cross-polarized image). For InSAR, only co-polarized image is necessary, while the cross-polarized data are to be included in future (only their intensity part). The SLC-C images are currently saved in uncompressed form (~280 MB/burst), while it was tested that a 7zip compression can minimize the size of one burst to ~80 MB. Table 1 below gives an overview of approximate size amounts of original files at the current date (3580 unique SLC images were in CollGS in October 2017) and with an expected value in December 2020 (7740 SLC images).

Table 1. Current and expected average data size of full Sentinel-1 SLC data in CollGS and co-polarized SLC-C data in IT4S1

dataset size	SLC (CollGS)*		SLC-C (IT4S1)	
	compr.	uncompr.	compr.	uncompr.
1 burst	200 MB	550 MB	80 MB	280 MB
current (Oct 2017)	16 TB	45 TB	6.5 TB**	23 TB**
expected (Dec 2020)	37 TB	102 TB	15 TB	52 TB

*including cross-polarized images

**predicted values for full coverage of the system

For ESD correction, the system merges all available bursts in a related swath overlapping Czechia within a current date and combines them with already preprocessed bursts. Thus, generation of one SLC-C image is performed for ~20 bursts at once. This takes approximately 24 core-hours on Salomon cluster. The expected computational load for SLC-C generation over all SLC data over Czechia is:

- all data until October 2017: 98 000 core-hours,
- data of Oct 2017 - Dec 2020: 125 000 core-hours.

The multitemporal InSAR processing itself is a (paradoxically) lower computational burden; depending on selected implementation, it is possible to count with few minutes (PS implementation using only temporal phase unwrapping) up to 24-48 core-hours per burst (PS with spatio-temporal unwrapping or 48-72 for SB techniques using a spatial filtering). An interferogram of one burst (combination of two SLC-C images) is generated within 8 seconds. It is expected that generation of PS result over the whole Czechia would take at least 92 000 core-hours, while it would be 180-280 thousands core-hours in the case of SB InSAR.

5. CONCLUSIONS AND FUTURE WORKS

While common HPC approaches of utilizing Sentinel-1 images for InSAR start their processing chain from original SLC data (e.g. ESA G-POD service [13]), the IT4S1 system allows a faster and more flexible multitemporal processing thanks to generation of pre-prepared SLC-C images. It is expected that the heavy computational load needed for nation-wide SLC-C data generation will be valorized in non-commercial applications for national geologic, urban planning, forestry or risk management spheres.

A large spatial coverage of processed data can lead to significant economic savings instead of e.g. installation of in-situ measurement tools such as GPS or another geodetic instruments for monitoring generally large areas of interest.

An automatic InSAR-based motion warning system can provide an early warning about a possibility of stability impact of critical infrastructure, based on analysis of changes in interferometric time series. A raised warning against landslide or displacements of a critical infrastructure can lead to significant economical savings. Presented IT4S1 framework is ready for further development in this direction. As additional functionality, the system is to store Sentinel-1 cross-polarization data for polarimetry analyzes – these can be used to identify e.g. deforestation or other changes in vegetation cover, while already stored radar intensity images can be used to identify significant structure changes, perform pixel offset tracking for evaluation of larger motions (e.g. landslides) etc.

6. ACKNOWLEDGMENTS

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project „IT4Innovations excellence in science - LQ1602“. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005), is greatly appreciated, as well as tools used for the system that is ISCE developed by Caltech and NASA/JPL, SARPROZ by Daniele Perissin (Purdue University) and STAMPS by Andrew Hooper (TU Leeds). Metadata database was based on scripts kindly offered by TU Leeds.

7. REFERENCES

- [1] Lazecký, M., M. Bakoň, I. Hlaváčová, J. J. Sousa, N. Real, D. Perissin, and G. Patricio, "Bridge Displacements Monitoring using Space-Borne SAR Interferometry", *Jour. of Sel. Top. in App. Earth Obs. and Rem. Sens. (J-STARS)*, IEEE, vol. PP, no.99, doi: 10.1109/JSTARS.2016.2587778, 2017.
- [2] Lazecký, M., M. Bakoň, D. Perissin, J. Papco, and S. Gamse, Analysis of dam displacements by spaceborne SAR interferometry, 9 pp., In *ICOLD 2017*, 2017.
- [3] F. C. Comut, A. Ustun, M. Lazecký, and D. Perissin, "Capability of detecting rapid subsidence with Cosmo SkyMed and

- Sentinel-1 dataset over Konya city”, 5 pp., In *ESA Living Planet Symposium*, ESA, ESA SP740, ISBN 978-92-9221-305-3, 2016.
- [4] Lazecký M., E. Jiráňková, and P. Kadlečík, “Multitemporal monitoring of Karvina subsidence trough using Sentinel-1 and TerraSAR-X interferometry”. *Acta Geodyn. Geomater.*, 14, No. 1 (185), 53–59, DOI: 10.13168/AGG.2016.0027, 2017.
- [5] Lazecky, M., F. C. Comut, E. Nikolaeva, M. Bakon, J. Papco, A. M. Ruiz-Armenteros, Y. Qin, J. J. Sousa, and P. Ondrejka, “Potential of Sentinel-1A for nation-wide routine updates of active landslide maps”, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLI-B7, 775-781, 2016.
- [6] Ferretti, A., C. Prati, and F. Rocca, “Nonlinear subsidence rate estimation using Permanent Scatterers in differential SAR interferometry”, *IEEE Trans. Geos. Rem. Sens.* 38 (5), 2000.
- [7] Berardino, P., G. Fornaro, R. Lanari and E. Sansosti, “A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms”, *IEEE Trans. Geosci. Remote Sens.* 40, 2375–2383, 2002.
- [8] Perissin, D., and T. Wang, "Repeat-pass SAR Interferometry with Partially Coherent Targets", *IEEE Transactions on Geoscience and Remote Sensing*, 50 (1), IEEE, pp. 271-280, 2012.
- [9] Yague-Martinez, N., P. Prats-Iraola, F. Gonzalez, R. Brcic, R. Shau, D. Geudtner, M. Eineder, and R. Bamler, “Interferometric Processing of Sentinel-1 TOPS Data”, *IEEE Trans. on Geos. and Rem. Sens.*, 54, pp. 1-15, 10.1109/TGRS.2015.2497902, 2016.
- [10] Z. Li, T. Wright, A. Hooper, P. Crippa, P. Gonzalez, R. Walters, J. Elliott, S. Ebmeier, E. Hatton, and B. Parsons, “Towards InSAR everywhere, all the time, with Sentinel-1”, *ISPRS Archives*, pp. 763–766, 2016.
- [11] Zebker, H.A., Hensley, S., Shanker, P., and Wortham, C. “Geodetically accurate InSAR data processor”, *IEEE Transactions on Geoscience and Remote Sensing* 48, 4309–4321, 2010.
- [12] Hooper, A., “A multi-temporal InSAR method incorporating both persistent scatterer and small baseline approaches”, *Geophysical Research Letters* 35, 2008.
- [13] F. Casu, S. Elefante, P. Imperatore, I. Zinno, M. Manunta, C. De Luca and R. Lanari, “SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation,” *IEEE JSTARS*, vol. 7, no. 8, pp. 3285-3296, 2014, doi: 10.1109/JSTARS.2014.2322671

THE VALUE OF SAR BIG DATA FOR GEOHAZARD APPLICATIONS: AUTOMATED GRID PROCESSING OF ERS-1/2 AND ENVISAT DATA IN ESA'S G-POD

F. Cigna & D. Tapete

Italian Space Agency (ASI), Via del Politecnico s.n.c., 00133 Rome, Italy

ABSTRACT

Long-term preservation of historical SAR data is crucial to guarantee the availability of these assets for geological hazard applications in the future. In this paper we provide examples from automated grid processing of long time series of ESA's ERS-1/2 and ENVISAT imagery using the Land Information hosted processing service 'InSAR SBAS' available through the G-POD platform. More than 1,000 medium resolution SAR scenes freely available through ESA's Virtual Archive 4 were used. Automated processing with the parallel SBAS (P-SBAS) chain allowed the estimation of ground stability and deformation across a number of areas of interest, including major cities in northern and southern Europe, Mexico and the Middle East. In these areas, natural hazards (e.g. ground settlement and volcanic activity) combine with anthropogenic factors (e.g. groundwater abstraction and engineering works). Associated ground motion velocities and time series were retrieved using P-SBAS with processing time demands of, on average, only half a day.

Index Terms— InSAR, big data, time series, ground motion, geological hazards

1. INTRODUCTION

Mapping, long-term monitoring and characterization of geological processes and hazards such as land subsidence, landslides and ground settlement with space-borne Interferometric Synthetic Aperture Radar (InSAR) require the availability of long stacks of SAR data spanning periods of a few months to up to several years.

Preserving SAR data archives and making them accessible, usable and exploitable for the scientific community and, more generally, the public is therefore crucial to guarantee that these digital assets and their scientific value will be available for geohazard applications and studies for the generations to come. This is particularly true for studies of global change and long-term trends and patterns, which require users to access time series of data spanning 25 or even more years.

This need for SAR data preservation is in line with the high level goals and objectives of the collaborative European Earth Observation (EO) Long Term Data Preservation Framework to jointly and cooperatively preserve space EO data from all European Space Agency (ESA) and Third Parties' ESA-managed missions (e.g. [1]).

In the current era of 'big SAR data', the volume and length of interferometric SAR data stacks are growing exponentially (e.g. every 6 days a new Sentinel-1 IW scene is acquired for all land areas of Europe, along each pass, ascending and descending), and with them InSAR processing workloads and demands. As a consequence, hardware and software requirements for advanced InSAR processing of long stacks of satellite SAR imagery to generate land deformation time series are notably increasing (e.g. [2]).

Today, a vast component of SAR data handling, initial manipulation and specialised InSAR processing can be delegated to remote systems and virtual environments. For instance, ESA's Grid-Processing On Demand (G-POD) platform for EO applications offers an environment where SAR data can be processed using high-performance and sizeable computing resources.

In this paper, a number of InSAR processing trials that were carried out using G-POD and its Land Information hosted processing service 'InSAR SBAS' will be presented.

2. INSAR TIME SERIES GENERATION IN G-POD

The Land Information Service 'InSAR SBAS' hosted in ESA's G-POD is based on the automated Parallel Small Baseline Subset (P-SBAS) processing chain developed at the Institute for Electromagnetic Sensing of the Environment of the National Research Council (IREA-CNR) of Italy [3-4]. The service allows both the generation of interferograms and the advanced multi-temporal analysis of time series, with extraction of full land deformation histories for large datasets of coherent targets.

In this work, more than 1,000 SAR scenes at medium resolution acquired by ESA's ERS-1/2 and ENVISAT missions and made freely available through ESA's Virtual Archive 4 were used to run several trials of automated InSAR processing of SAR big data stacks to generate derived products with value for geohazard applications.

Processing with P-SBAS was conducted through the user-friendly G-POD web portal, which allowed selection of input data, setting of processing parameters, thresholds and options, and effective monitoring of the full processing chain from remote, as well as downloading of the generated results for subsequent visualisation in Google Earth and uploading into GIS platforms for interpretation and analysis.

Case studies that were analysed include major cities of northern and southern Europe such as Naples and London, as

well as in Mexico and the Middle-East, where natural hazards such as land subsidence, ground settlement, seismicity and volcanic activity, combine with anthropogenic factors, such as groundwater abstraction and engineering works (Fig.1).

Whereas for a number of areas the P-SBAS results showed a general stability of the analysed cities, in a number of locations the observed ground deformation velocity fields revealed sectors affected by land instability with rates as high as some tens of mm per year.

In the city of Rome (Italy), for instance, both the ERS-1/2 and ENVISAT based analyses for the periods 1992-2000 and 2002-2010 (Fig.2) confirmed the presence of movements that occurred in the direction away from the satellite sensor (negative velocities), i.e. land subsidence. The latter indicates

the occurrence of natural compaction of alluvial deposits with annual velocity of up to 10-15 mm/year is clearly depicted along the Tiber River, a phenomenon well known and documented in the literature, e.g. [5-6].

Similarly, patterns of land uplift were retrieved in 2002-2010 in the area of the Phlegrean Fields, west of the city of Naples (Fig.1a), ground settlement in central London (Fig.1b), Mexico City (Fig.1c) and several towns in the Middle East (Fig.1d) as an effect of groundwater pumping and compaction of the local aquifer systems. As in the case of Rome, these areas have been historically affected by geological processes and land instability, often influenced by human activities, as observed by other published studies using satellite InSAR methods and data, e.g. [7-9].

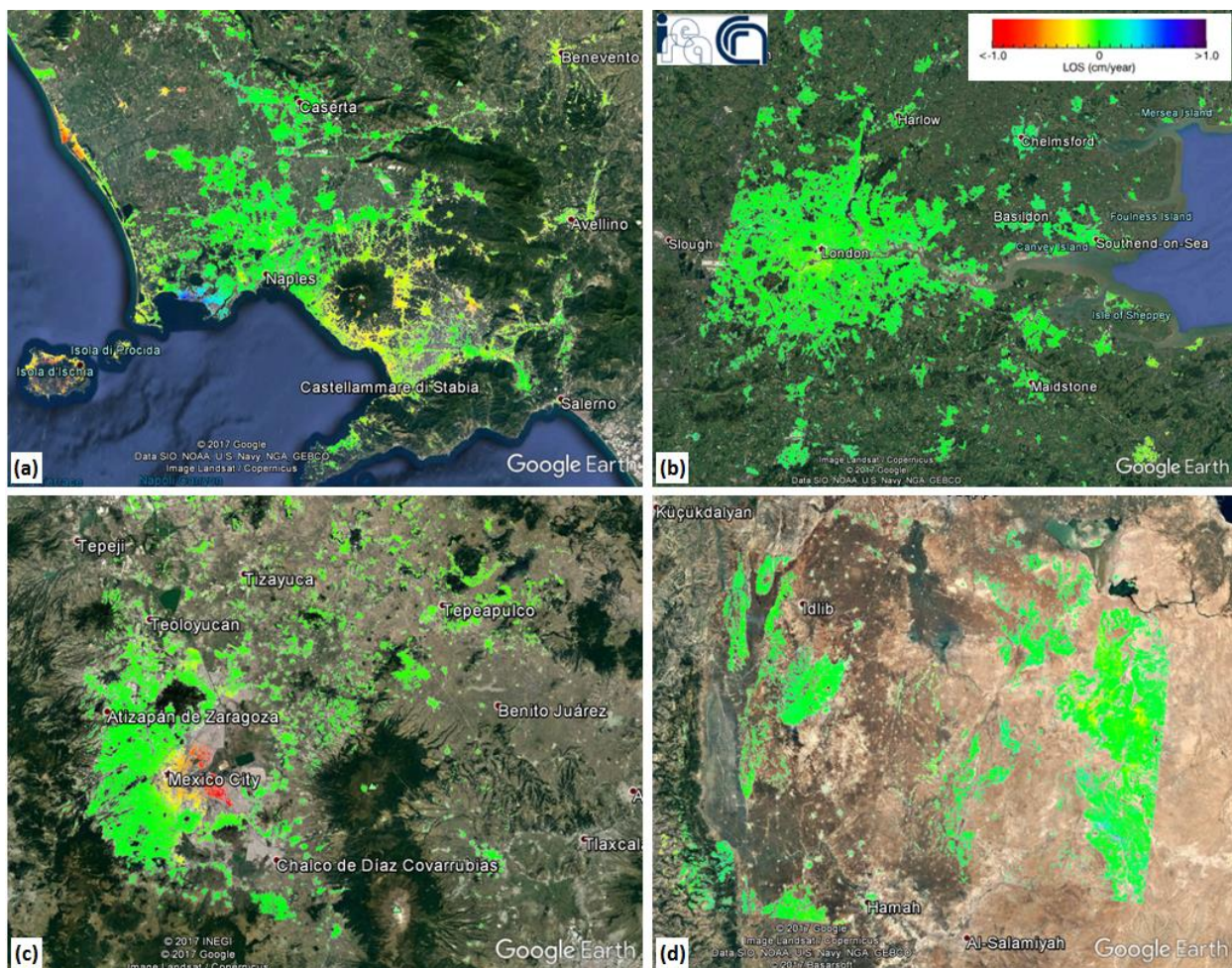


Fig.1: Satellite InSAR results for the test areas of (a) Naples (Italy), (b) London (UK), (c) Mexico City (Mexico) and (d) Hamah (Syria), obtained after P-SBAS processing of ENVISAT data using G-POD. The results are displayed according to the observed annual ground deformation. Negative velocities (yellow to red targets) indicate areas exhibiting motion away from the satellite sensor (e.g. subsidence in Mexico City), while positive velocities (light blue to violet) indicate areas affected by motion towards the sensor (e.g. uplift observed in the Phlegrean Fields).

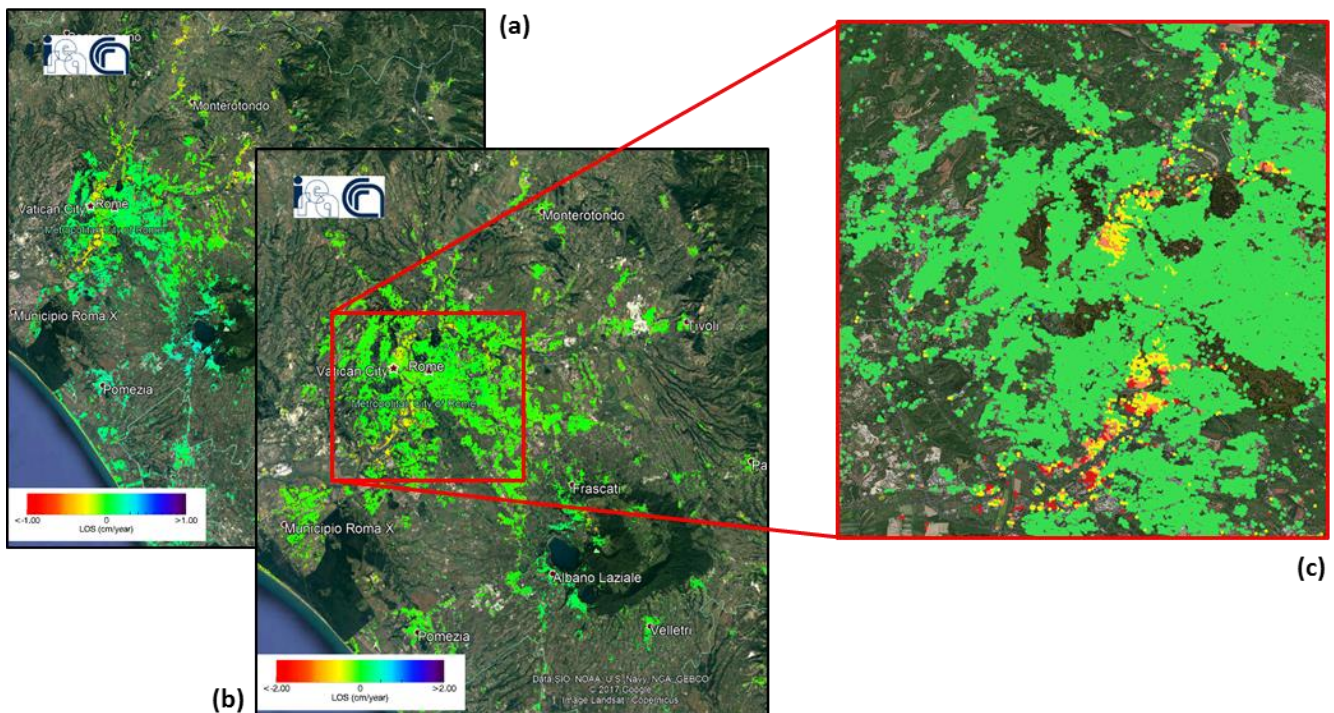


Fig.2: P-SBAS results for the urban area of Rome (Italy) based on (a) 45 ENVISAT scenes in descending mode (2002-2010) and (b-c) 66 ERS-1/2 scenes in descending mode (1992-2001). While most of the city appears stable based on the P-SBAS results, natural compaction of alluvial deposits with annual velocity of up to 10-15 mm/year is clearly depicted along the Tiber River.

We performed an analysis of processing times needed to derive time series for each case study, based on a set of 33 trials, all using the same Computing Element: *ESA CE 01 SL6 64bits*, and an average number of input satellite ERS-1/2 or ENVISAT SAR scenes of 30.

Total processing time (from submission of the trial to publication of the results) were, on average, of only half a day. By plotting time demands against the input number of SAR scenes processed in each trial (Fig.3), we also observed a good correlation and, via linear regression, derived a simple empirical relationship: $T = 0.38 * N_{SAR}$, where T is the total processing time in hours, and N_{SAR} is the number of input ERS or ENVISAT scenes used. Similarly, by analysing time demands against the number of small baseline interferograms processed in each trial (N_{inf}), we derived: $T = 0.13 * N_{inf}$.

Based on these trials, three factors are worth specific consideration:

- (1) the extremely short time demands that we observed;
- (2) the precision (up to mm) of the retrieved results;
- (3) the added value of their geological interpretation.

Open platforms and tools, such as the P-SBAS in G-POD, allow demanding InSAR workloads to be handled, regardless of the operators' background and accessibility to specialist software and computing facilities. Having the processing component solved, the user community can then focus on

generating InSAR products from open SAR data to gain extremely important knowledge about land dynamics and processes.

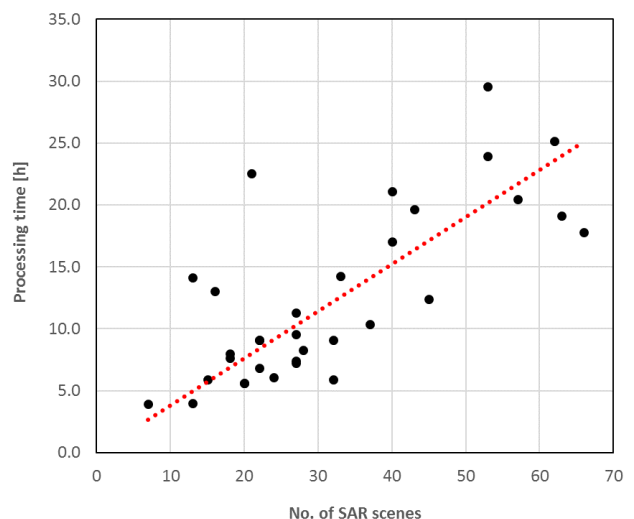


Fig.3: Observed processing time demands for the P-SBAS chain to complete the advanced multi-temporal processing of ERS and ENVISAT stacks, based on 33 trials using *ESA CE 01 SL6 64bits* computing element in G-POD.

4. CONCLUSIONS

Our trials with ESA's G-POD are examples of the wide spectrum of geological applications that users can cover with G-POD.

In this context, the value of InSAR big data is manifold: they serve a diverse scientific community which will increase as more archive data are made available and more geographic regions across the world are covered; they are provided with extremely reduced processing time, thereby allowing InSAR expert users to concentrate on adding value to the processing outputs; they are offered as standardised and reliable products so even non-expert users can use them making their studies adherent to state-of-the-art EO techniques.

Availability of historical data is crucial to sense the dynamics of geological processes that can develop over many years and may be characterised by rates of a few to several mm/year (not always visible to the naked eye), affect infrastructure, quarters or even entire cities or regions, and/or change deformation trend with time.

5. REFERENCES

- [1] V. Beruti, and M. Albani, "The ESA Earth Observation Long Term Data Preservation (LTDP) Programme," *PV 2009 Conference: Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*, 1-3 Dec 2009, ESAC, Villafranca, Spain, pp. 8, 2009.
- [2] F. Cigna, "Getting ready for the generation of a nationwide ground motion product for Great Britain using SAR data stacks: feasibility, data volumes and perspectives," *Proc. of 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1464-1467, 2015.
- [3] F. Casu, S. Elefante, P. Imperatore, I. Zinno, M. Manunta, C. De Luca, and R. Lanari, "SBAS DInSAR parallel processing for deformation time series computation," *IEEE JSTARS*, 7(8), pp. 3285-3296, 2014.
- [4] C. De Luca, R. Cuccu, S. Elefante, I. Zinno, M. Manunta, V. Casola, G. Rivolta, R. Lanari, and F. Casu, "An On-Demand Web Tool for the Unsupervised Retrieval of Earth's Surface Deformation from SAR Data: The P-SBAS Service within the ESA G-POD Environment," *Remote Sens.*, 7, 15630-15650, 2015.
- [5] F. Cigna, R. Lasaponara, N. Masini, P. Milillo, and D. Tapete, "Persistent Scatterer Interferometry Processing of COSMO-SkyMed StripMap HIMAGE Time Series to Depict Deformation of the Historic Centre of Rome, Italy," *Remote Sensing* 6(12), 12593-12618, 2014.
- [6] D. Tapete, R. Fanti, R. Cecchi, P. Petrangeli, and N. Casagli, "Satellite radar interferometry for monitoring and early-stage warning of structural instability in archaeological sites," *Journal of Geophysics and Engineering* 9, S10-S25, 2012.
- [7] A. Sowter, M. Che Amat, F. Cigna, S. Marsh, A. Athab, and L. Alshammari, "Mexico City land subsidence in 2014-2015 with Sentinel-1 IW TOPS: first results using the Intermittent SBAS (ISBAS) technique," *Int. J. of Applied EO and Geoinformation*, 52, 230-242, 2016.
- [8] R. Lanari, O. Mora, M. Manunta, J.J. Mallorquí, P. Berardino, and E. Sansosti, "A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms," *IEEE Trans. Geosci. Remote Sens.* 42, 1377-1386, 2004.
- [9] D. Tapete, F. Cigna, A. Sowter, and S. Marsh, "Small Baseline Subset (SBAS) pixel density vs. geology and land use in semi-arid regions in Syria," *Proc. of 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3353-3356, 2015.

EXPLOITING OCEAN OBSERVATION AND SIMULATION BIG DATA TO IMPROVE SATELLITE-DERIVED GEOPHYSICAL PRODUCTS: ANALOG STRATEGIES

R. Fablet¹, P. Viet¹, R. Lguensat¹, P.H. Horrein¹, B. Chapron²

(1) IMT Atlantique; Lab-STICC, Brest, France, (2) Ifremer; LOPS, Brest, France

ABSTRACT

The ever increasing geophysical data streams pouring from earth observation satellite missions and numerical simulations along with the development of dedicated big data infrastructure advocate for truly exploiting the potential of these datasets, through novel data-driven strategies, to deliver enhanced satellite-derived geophysical products from partial satellite observations. We here demonstrate a proof-of-concept of the analog data assimilation for an application to the reconstruction of cloud-free level-4 gridded Sea Surface geophysical fields, namely Sea Surface Temperature (SST) and Sea Surface Height (SSH). Our results point out the relevance of big-data-oriented analog strategies to benefit from large-scale observation and/or simulation datasets for enhanced satellite-derived geophysical products.

Index Terms— Data assimilation, data-driven & analog methods, distributed computing, space-time interpolation, L4 satellite-derived products

1. PROBLEM STATEMENT AND RELATED WORK

The delivery of level-4 gridded geophysical products is a key issue for operational oceanography and meteorology, as most often in situ and/or satellite-derived observations involve irregular sampling patterns, which relate to satellite orbits as well as to the sensitivity of the sensors to the atmospheric conditions. Satellite-derived observation of the sea surface temperature (SST) from infrared sensors is a typical example. It may involve very large cloud-related missing data rates (>90%), which makes the spatio-temporal interpolation critical. Whereas operational level-4 products mainly resort to optimal interpolation techniques, using some prior on the space-time covariance of the geophysical field of interest [1], the availability of large-scale historical observation and/or simulation datasets advocate for the exploration of fully data-driven and big-data-oriented strategies.

Data-driven strategies have emerged in the field of signal and image processing as particularly appealing and efficient strategies for solving signal and image inverse problems [2, 3, 4]. The key idea is that large-scale datasets provide an implicit representation of the underlying variability of the process of interest. A number of models and algorithms

have been proposed to turn such an implicit representation into a computationally-efficient strategy, including for instance learning-based and analog strategies [2, 3, 5]. In this context, we recently introduced the analog data assimilation [6], which provides an implementation of such a data-driven framework for dynamical systems. The key component is to build a dynamical model to simulate new state dynamics from the analogy between a current state and the state dynamics previously observed and/or simulated. We demonstrated in [6] the relevance of the analog data assimilation for low-dimensional chaotic state dynamics.

In this study, we further address such strategies for the reconstruction of time series of satellite-derived geophysical fields. The search for analogs to some current state faces the curse of dimensionality and fails for high-dimensional state [7]. Though it is unrealistic to collect a sufficiently large observation dataset, such that at a global or regional scale a current geophysical state matches any of the previously observed states at both the regional/global scale and all fine scales, we argue that one can propose to decompose the considered high-dimensional problem as a series of low-dimensional problems which analog data assimilation efficiently applies to from available large-scale datasets. Using scale-space decomposition principles [8], we demonstrate the relevance of this strategy for two case-studies, namely the reconstruction of sea surface temperature fields from cloudy satellite observations and the reconstruction of sea surface anomaly from along-track satellite data. We evaluate its big-data-oriented implementation, including a comparison to state-of-the-art model-driven and data-driven spatio-temporal interpolation techniques.

This paper is organized as follows. Section 2 introduces the proposed analog assimilation model. Section 3 presents experimental results for an application to missing data interpolation associated with satellite-derived sampling patterns.

2. PROPOSED ANALOG FRAMEWORK

Formally, we follow the state-space setting classically considered for stochastic data assimilation [9] :

$$\begin{cases} \mathbf{x}(t) &= \mathcal{M}(\mathbf{x}(t-1)) \\ \mathbf{y}(t) &= \mathcal{H}(\mathbf{x}(t), \Omega(t)) + \eta \end{cases} \quad (1)$$

where t is a discrete time index, \mathbf{x} the hidden state sequence to be reconstructed and \mathbf{y} the observed data sequence. \mathcal{M} is the dynamic prior on state $\mathbf{x}(t)$ given previous state $\mathbf{x}(t-1)$. Θ refers to the parameterization of model \mathcal{M} . \mathcal{H} is the observation operator, where $\Omega(t)$ is the missing data mask at time t . η is a random process accounting for uncertainties.

The proposed scheme involves three key components :

- **A multiscale decomposition :** At any time t , the field $\mathbf{x}(t)$ is decomposed as a sum of a large-scale component $\bar{\mathbf{x}}$ and of a fine-scale component $d\mathbf{x}$: $\mathbf{x}(t) = \bar{\mathbf{x}} + d\mathbf{x}(t)$. The large-scale component typically accounts for scales above 100km and the fine-scale one for scales below 100km.

- **A patch-based decomposition of the fine-scale component :** following patch-based image representations [2], we decompose image $d\mathbf{x}(t)$ into overlapping $K \times K$ patches (typically, $K = 40$) and apply an EOF-based decomposition to each patch : $d\mathbf{x}(t)(\mathcal{P}_s) = \sum_{k=1}^{N_{EOF}} \alpha_{t,s,k} B_k$, where B_k are the EOFs and $\alpha_{t,s,k}$ the coefficients of the decomposition of patch $d\mathbf{x}(t)(\mathcal{P}_s)$ onto the EOFs.

- **Scale-specific dynamic models :** we consider different dynamic models for the large-scale and fine-scale components. As the large-scale involves a spatial smoothing over a large domain, a Gaussian prior associated with an optimal interpolation model naturally applies [1]. Besides, the patch-based representation results in a local low-dimensional representation of the fine-scale component within each patch, which the analog data assimilation [6] applies to. We let the reader refer to [6] for details on the considered analog dynamic models, namely locally-incremental and locally-linear models. In both cases, they result in first retrieving analogs to a current state and estimating from these analogs and their successors the dynamical transform to be applied to the current state.

The numerical resolution of our analog model comes first to solving for the large-scale component using an optimal interpolation and second to independently solving for the reconstruction of the fine-scale component for each patch using an ensemble Kalman smoother with the considered analog priors [6]. The last step spatially averages the reconstruction of the fine-scale component issued for each patch location.

3. APPLICATION TO SEA SURFACE GEOPHYSICAL FIELDS

We design numerical experiments to evaluate the performance and the computational complexity of the proposed strategy for the spatio-temporal interpolation of sea surface geophysical fields from partial observations. We consider two case-studies : SST (Sea Surface Temperature) and SSH (Sea Surface Temperature). Whereas satellite-derived SST data involve large missing data rate for high-resolution infrared sensors due to their sensitivity to the cloud cover, satellite altimeters are narrow-swath sensors, which result in a scarce sampling of the sea surface. In both cases, we implement an Observing System Simulation Experiment (OSSE) to evaluate

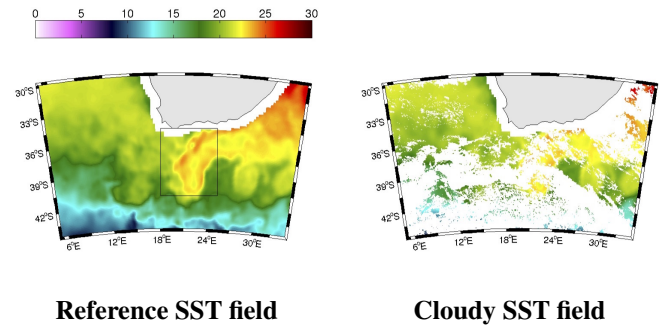


Fig. 1. Sampling pattern associated with cloudy SST data : groundtruthed AMSR-E SST field at a given date (left), simulated cloudy SST field generated using METOP SST cloud mask (right).

quantitatively and qualitatively the proposed analog strategies. We detail below the implemented case-studies.

All experiments were performed on Teralab big data platform (<https://www.teralab-datascience.fr>) with the following multi-core configuration : 24 virtual CPUs with a 64G RAM. The Python code of the considered analog data assimilation scheme is available from https://github.com/rfablet/PB_ANDA/.

3.1. SST case-study

We use a reference cloud-free SST time series [1] to build a representative groundtruthed dataset for METOP cloud patterns off South Africa. We consider a 8-year daily time series of 600×300 images, *i.e.* a $\sim 150\text{-}000$ -dimensional state-space when removing land pixels. We use the first seven years as an historical dataset to retrieve analogs and apply the proposed analog data assimilation to the last year.

Reported results (Tab.1) demonstrate the relevance of the proposed analog strategy. While, as expected, the straightforward application of the analog data assimilation to the regional scale (G-AnEnKS) does not bring a significant improvement due dimensionality issues compared to the large-scale optimal interpolation (OI), we report significant improvement for the reconstruction of the fine-scale component using the proposed multiscale strategy compared to both OI and state-of-the-art EOF-based interpolation [10]. Visual comparisons of interpolation results may be found in [11] along with the spectral analysis of the reconstructed fields.

Regarding its big-data-oriented implementation, additional experiments support the good parallelization performance of the proposed strategy based on the independent patch-level processing for the fine-scale component. In this respect, in addition to efficient analog search schemes, a significant complexity gain with no significant loss in reconstruction performance can be retrieved for clustered locally-linear analog operators (cf. PB-AnEnKS(C-LL) model in Tab.1). The later typically leads to a reduction of the computational complexity

of the locally-linear analog forecasting by a factor of 3.

Table 1. Reconstruction performance for the SST case-study : optimal interpolation (OI), global (G-AnEnKS) and patch-based (PB-AnEnKS) analog assimilation with locally-incremental (LI), locally-linear (LL) and clustered locally-linear (C-LL). We report the mean relative Mean Square Error (RMSE) and the correlation coefficient (ρ) for a 1-year daily SST time series. We let the reader refer to the main text for the details of considered experimental setting.

Method	RMSE	ρ
OI	0.49 ± 0.06	0.66 ± 0.06
G-AnEnKS	0.43 ± 0.05	0.72 ± 0.04
PB-DinEOF	0.44 ± 0.05	0.72 ± 0.04
PB-AnEnKS (LI)	0.30 ± 0.06	0.87 ± 0.03
PB-AnEnKS (LL)	0.27 ± 0.04	0.90 ± 0.03
PB-AnEnKS (C-LL)	0.27 ± 0.04	0.90 ± 0.03

Table 2. Reconstruction performance for the SSH case-study : optimal interpolation (OI), patch-based DinEOF interpolation (PB-DinEOF), global (G-AnEnKS) and patch-based (PB-AnEnKS), using a locally-linear analog forecasting model. We report the mean relative Mean Square Error (RMSE) and the correlation coefficient (ρ) for a 1-year 3-daily SLA (Sea Level Anomaly) time series. We let the reader refer to the main text for the details of considered experimental setting.

Method	RMSE	ρ
OI	0.026 ± 0.007	0.81 ± 0.08
G-AnEnKS	0.020 ± 0.006	0.89 ± 0.04
PN-DINEOF	0.023 ± 0.007	0.85 ± 0.07
PB-AnEnKS	0.013 ± 0.005	0.96 ± 0.02

3.2. SSH case-study

For the SSH OSSE, we consider a four-altimeter spatio-temporal sampling configuration using along-track data positions from Jason-2, Cyrosat-2, Saral-Altika and Hy-2A missions in 2014. Using 50-year 3-daily OGCM (Ocean General Circulation Model) data as groundtruthed reference, we simulate daily along-track data from the OGCM SLA data (Sea Level Anomaly). We apply the analog data assimilation for a 3-daily time step. We use the 49 first years of OGCM data as training data and we evaluate interpolation performance for the last year. The study zone is selected in the

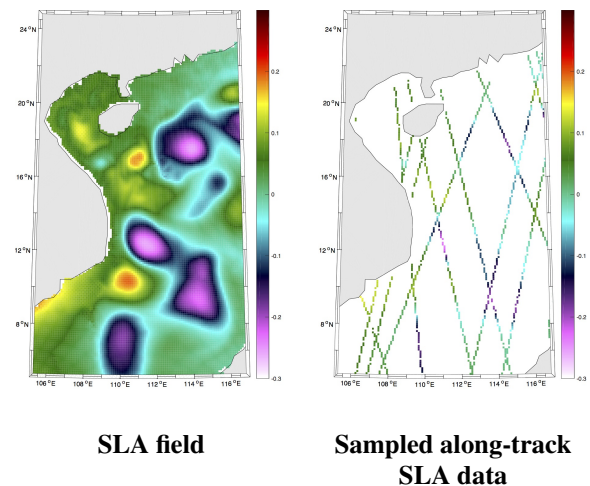


Fig. 2. Sampling pattern associated with along-track altimeter data : groundtruthed Sea Level Anomaly (SLA) field at a given date (left), sampled along-track altimeter data for the same date using a dataset of real along-track positions from four satellites, namely Jason-2, Cyrosat-2, Saral-Altika and Hy-2A missions (right).

South China Sea ($105^{\circ}\text{E}-117^{\circ}\text{E}$, $5^{\circ}\text{N}-25^{\circ}\text{N}$). We illustrate in Fig.2 an example of reference SLA field and the associated along-track SLA data simulated from real along-track positions. Here we simulated noise-free altimeter data.

The experimental setting of the patch-based analog data assimilation is defined as follows : 20×20 patches with 15-dimensional EOF decompositions, which typically accounts for 99% of the data variance for the considered dataset. As low-resolution component, we consider an optimal interpolation [12] with a spatial correlation length of 100km and a temporal correlation length of 15 days. Similarly to the SST case-study, we compare the reconstruction of the patch-based analog data assimilation to an optimal interpolation [12], an EOF-based interpolation (PB-DinEOF) and a direct application of the analog data assimilation at the regional scale (G-AnEnKS). For this second case-study, we only consider the patch-based analog assimilation with locally-linear analog forecasting models.

We report the mean interpolation performance in Tab.2. Similarly to the SST case-study, the analog data assimilation leads to a significant improvement of the interpolation performance. It outperforms by about 50% and 40% the optimal interpolation and DinEOF interpolation. It may be noted that the direct application of the analog data assimilation to the regional scale using an EOF decomposition over the entire region (G-AnEnKS model) leads to a significantly lower improvement of the reconstruction (namely, a gain of 13% w.r.t. OI for G-AnEnKS, and of 50 % for PB-AnEnKS). This is further illustrated visually in terms of retrieval of finer structures using PB-AnEnKS. Similarly conclusions can be drawn

when considering noisy along-track altimetry datasets. We let the reader to [13] for additional experiments and discussion of the key features of the proposed framework for the reconstruction of sea surface dynamics.

4. CONCLUSION

Through two different case-studies, we demonstrated in this study the relevance of analog strategies for the reconstruction of sea surface geophysical fields from partial observations due to the sampling pattern of narrow-swath satellites and to the sensitivity of satellite sensors to the atmospheric conditions. Given a reference groundtruthed dataset, we state the reconstruction issue as a data assimilation problem, where the dynamical model relies on analog forecasting operators. We reported potential gain up to 40-50% in terms of root mean square error with respect to an optimal interpolation, which is the classical interpolation solution for satellite-derived geophysical products. A key feature of the proposed analog framework is the decomposition of the global assimilation problem into a series of local patch-level assimilation issues, which benefit from a low-dimensional EOF-based representation of patch-level variability. By nature, such a patch-level architecture scales up to large-scale dataset linearly with respect to the number of patches.

Overall, these results open new avenues for the operational use of analog strategies for the reconstruction of higher-resolution geophysical products. A great interest of such analog strategies is also their ability to exploit multi-source synergies [14] either through the definition of a multi-source kernel to retrieve analogs or through the use of multi-source regression variables in the locally-linear analog forecasting model. From an operational perspective, analog data assimilation may exploit large-scale observation datasets and/or large-scale simulation datasets. Future work should further investigate the sensitivity to the size and representativeness of the considered training datasets. This appears to be critical for operational applications.

Acknowledgments : This work was supported by ANR (Agence Nationale de la Recherche, grant ANR-13-MONU-0014), Labex Cominlabs (grant SEACS), Teralab (grant TIAMSEA) and CNES (grant OSTST-MANATEE).

5. REFERENCES

- [1] C. J. Donlon, M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Xindong, "The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system," *Remote Sensing of Environment*, vol. 116, pp. 140–158, Jan. 2012.
- [2] G. Peyré, S. Bogleux, and L.D. Cohen, "Non-local Regularization of Inverse Problems," *Inverse Problems and Imaging*, vol. 5, no. 2, pp. 511–530, 2011.
- [3] R. Fablet and F. Rousseau, "Missing data super-resolution using non-local and statistical priors," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept. 2015, pp. 676–680.
- [4] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video Inpainting of Complex Scenes," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, Jan. 2014.
- [5] L. He, R. Fablet, B. Chapron, and J. Tournadre, "Learning-Based Emulation of Sea Surface Wind Fields From Numerical Model Outputs and SAR Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 10, pp. 4742–4750, Oct. 2015.
- [6] R. Lguensat, P. Tandeo, P. Aillot, and R. Fablet, "The Analog Data Assimilation," *Monthly Weather Review*, 2017.
- [7] H. M. Van Den Dool, "Searching for analogues, how long must we wait?," *Tellus A*, vol. 46, no. 3, 1994.
- [8] S. Mallat, *A wavelet tour of signal processing, second edition*, Academic Press, 1999.
- [9] G. Evensen, *Data Assimilation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [10] B. Ping, F. Su, and Y. Meng, "An Improved DINEOF Algorithm for Filling Missing Values in Spatio-Temporal Sea Surface Temperature Data," *PLOS ONE*, vol. 11, no. 5, pp. e0155928, May 2016.
- [11] R. Fablet, P. H. Viet, and R. Lguensat, "Data-driven Models for the Spatio-Temporal Interpolation of satellite-derived SST Fields," *IEEE Transactions on Computational Imaging*, 2017.
- [12] R. Escudier, J. Bouffard, A. Pascual, P.-M. Poulain, and M.-I. Pujol, "Improvement of coastal and mesoscale observation from space : Application to the northwestern Mediterranean Sea," *Geophysical Research Letters*, vol. 40, no. 10, pp. 2148–2153, 2013.
- [13] R. Lguensat, P. Huynh Viet, M. Sun, G. Chen, T. Fenglin, B. Chapron, and R. Fablet, "Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation," Tech. Rep., Oct. 2017.
- [14] R. Fablet, J. Verron, B. Mourre, B. Chapron, and A. Pascual, "Improving mesoscale altimetric data from a multi-tracer convolutional processing of standard satellite-derived products," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

SENTINEL-2 DASHBOARD FOR SPATIO-TEMPORAL ANALYSIS OF GLOBAL SCENE COVERAGE

Martin Sudmanns, Hannah Augustin, Anna-Maria Cavallaro, Dirk Tiede, Stefan Lang

Department of Geoinformatics – Z_GIS, University of Salzburg, Austria

ABSTRACT

The implemented Sentinel-2 Dashboard deals with metadata of all accessible Sentinel-2 scenes, allowing users to geospatially analyse Sentinel-2 metadata using geovisualisation in an exploratory manner. It enables the investigation and visualisation of Sentinel-2 scenes' inherent spatio-temporal dynamics at a global scale. Examples include granule maps with aggregated statistics, per-granule statistics, and current and predicted Sentinel-2 positions. The Sentinel-2 Dashboard can be seen as an initial big data filtering and evaluation system. To the best of our knowledge, there is currently no other freely available tool with similar features.

Index Terms— Sentinel-2, Metadata, Scene Coverage, Big Earth Data

1. INTRODUCTION AND MOTIVATION

The Sentinel-2 satellites acquire terabytes of data on a daily basis, which are stored and made available to users by various access portals. The technical specifications are conveyed to users as being simple: pixel size of 10m, 20m or 60m depending on the spectral band, 290 km swath size, revisit time of 5 days at the equator, etc [1]. Investigations of the theoretical coverage of Sentinel-2A are also available [3]. However, the actual spatio-temporal coverage is more complex. Differing levels of cloudiness across the Earth, the geographic location and the acquisition plan all affect the data availability, suitability and quality. For example, a time series analysis in Europe operates on data with different temporal characteristics than an analysis in tropical regions or deserts in Australia. Analyses on global to regional scales usually do not take the spatio-temporal inhomogeneity of data coverage into account.

A systematic analysis and evaluation of data availability and data quality reveals the inherent spatio-temporal dynamics of the captured Sentinel-2 scenes and allows drawing conclusions about their suitability for a specific application. Such an analysis enables initial upstream big data spatio-temporal filtering and a better estimation of the validity of planned analysis. Interpretation of analysis results at global to regional scales with study areas larger than a single granule may benefit from taking the inhomogeneity of the underlying

data into account. Apart from technical and scientific use, geovisualisation of the spatio-temporal dynamics can also be used to communicate to the public in a more appealing way just how massive the amount of Sentinel-2 data are.

In the presented approach, the Sentinel-2 metadata are systematically harvested and visualised using an implemented Web application tool called the *Sentinel-2 Dashboard*. In addition to spatio-temporal exploration of metadata, the Sentinel-2 Dashboard allows investigating the satellite status, including satellite positions, predicted orbits, mission reports or data quality reports, and links to publicly available third-party information about Sentinel-2. This makes the Sentinel-2 Dashboard a unique solution and one-stop shop for Sentinel-2 related information.

2. SENTINEL-2 METADATA USED IN THE DASHBOARD

The Sentinel-2 Dashboard uses all of the accessible Sentinel-2 metadata available from the Copernicus Open Access Hub with on-going harvesting, including level 1C- and 2A products. The metadata variables that are currently used are the granule-footprint, cloud cover percentage, acquisition time and processing time. Based on the information about the number of scenes the following information is produced per granule-footprint: average cloud cover; percentage of scenes with less than 10% cloud cover; average time until ingestion; time between acquisitions; and average time between cloud-free acquisitions.

Data about the orbit parameters to calculate the current positions of the Sentinel-2 satellites and to predict the next orbits are collected from the CelesTrak website [2]. Publicly available information about Sentinel-2 is either collected from the ESA website (e.g. acquisition plans) or linked to it (e.g. Sentinel-2 news, data quality reports, mission reports).

3. SENTINEL-2 DASHBOARD

The Sentinel-2 Dashboard allows the user to: (1) create interactive granule-maps based on the aforementioned variables; (2) retrieve granule-based statistics using an infographics-style; (3) view the status of Sentinel-2A and -2B satellites on a map (e.g. current position, predicted orbits, acquisition plans); and (4) take a tour, which guides users to pre-defined interesting locations (e.g. with salient

acquisition characteristics). Additionally, an updated weather forecast is under development, which provides information about the temperature, precipitation, cloud cover, visibility and the solar radiation of the upcoming acquisition areas. Having this functionality, the Sentinel-2 Dashboard is meant to provide: (1) a tool to investigate the inherent spatio-temporal dynamics of Sentinel-2 imagery and the performance of the Sentinel-2 satellites; (2) a planning tool to examine and compare study areas with respect to their data availability and characteristics; (3) a possible validation tool of these characteristics, especially the cloud cover estimations reported in the metadata, e.g., by the planned weather widget; (4) detection of long-term changes and trends, e.g. seasonality of cloud cover; and (5) means for dissemination and communication of information about the Sentinel-2 mission and acquired imagery for improved education of students and the public.

One possible application use for the dashboard is as a meta-tool for the increasingly upcoming big data technology initiatives, including DIAS, and more broadly in all areas located in the Sentinel-2 high-resolution domain. For example, users might be interested in monitoring deforestation in Indonesia based on change detection using a Sentinel-2 time-series. Using the Sentinel-2 Dashboard they can select the relevant image footprints (granules) and immediately get relevant information about the frequency of observations, historic metadata of the average time between cloud-free acquisitions etc. This information in combination with the implemented real-time weather forecast guarantees a validation of the statistics and a solid basement to plan further applications with Sentinel-2 scenes and helps in

applying appropriate methodologies for the analyses. The information is made available in the web application with only a few interactions and within a very short amount of time (near-real time).

The technology of the Sentinel-2 Dashboard is developed in-house. The user frontend is an Ember application, which uses additional libraries for generating maps (leaflet) and handling time (moment). It accesses a REST API, which is based on python Flask. The data is stored in PostgreSQL with the spatial PostGIS extension enabled.

4. PRELIMINARY RESULTS

The first version of the Sentinel-2 Dashboard has already been developed and is operated by the Department of Geoinformatics – Z_GIS at the University of Salzburg. It will be made publicly available at the beginning of 2018. As of mid-October 2017, the metadata database encompasses data for approx. 2 million Sentinel-2A scenes and provides four different views or modules for users.

One view is based on granule-maps that are generated by mapping metadata statistics to each granule-footprint (Figure 1). The currently available statistics are: number of available scenes, average cloud cover, percentage of cloud-free scenes and average duration between cloud-free acquisitions. In these cases, cloud-free is defined as having less than 10 % clouds based on the cloud cover estimations reported in the metadata.

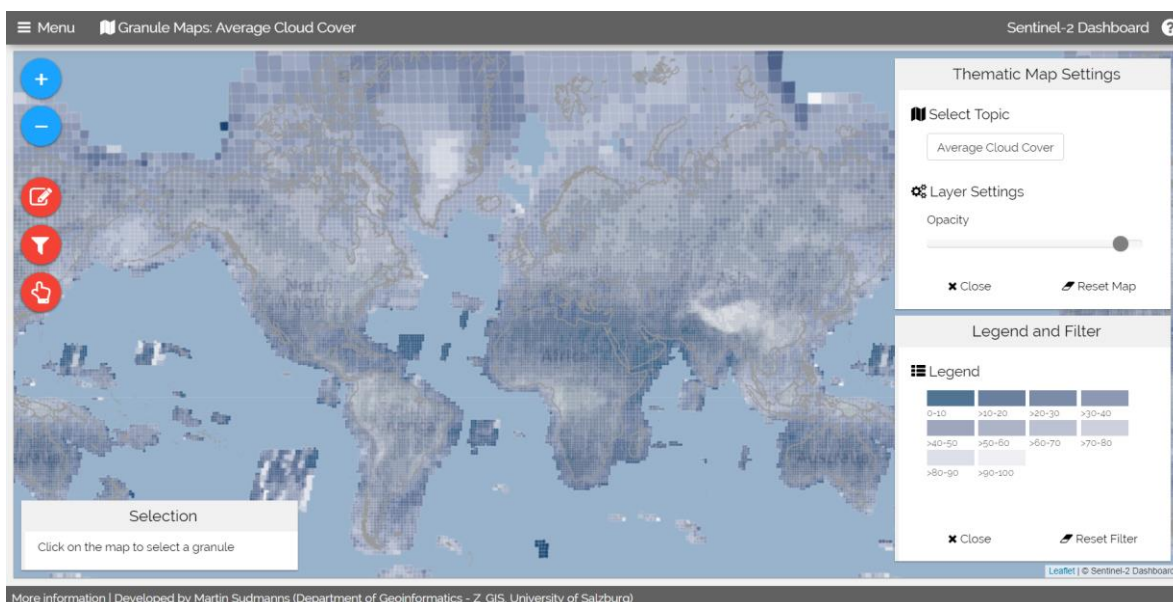


Figure 1: Map showing the global average cloud cover per granule, the darker the blue, the lower the cloud cover.

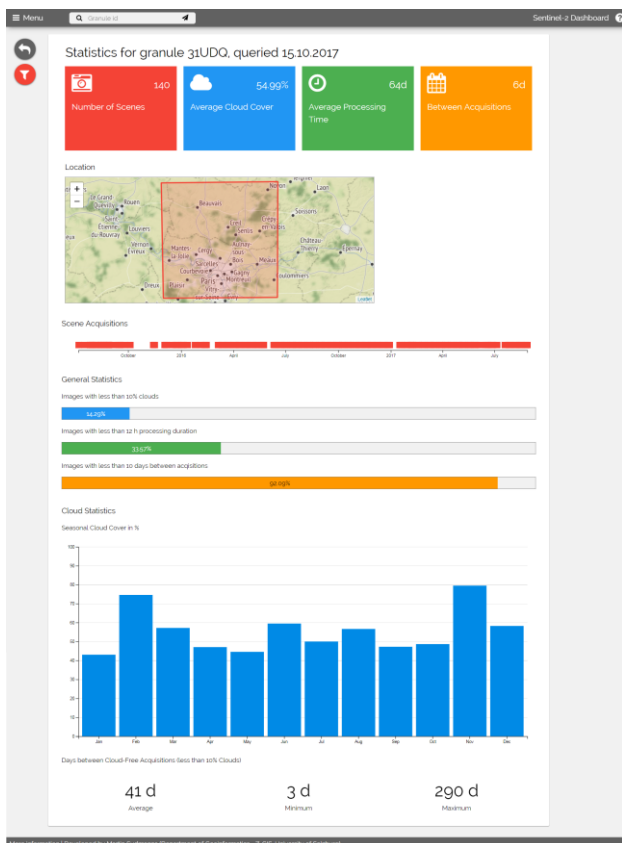


Figure 2: The Sentinel-2 Dashboard allows querying of detailed metadata statistics per granule.

The second view utilises an infographics style to present detailed, temporal statistics for individual granules (Figure 2). The available variables are the number of scenes, average cloud cover, processing duration, duration between acquisitions, date of the historic acquisitions and statistics about the percentage of images with less than 10% clouds, less than 12 h processing duration, and less than 10 days between acquisitions. Further, a bar chart allows investigation of the average cloud cover per month and the average, minimum and maximum days between two cloud-free acquisitions.

A third module allows investigation of the current positions and predicted orbits of the Sentinel-2 satellites together with the current acquisition plan and reports provided by ESA about the mission status and data quality (Figure 3).

The last module is educational and uses a story map to guide users step-by-step to pre-selected regions that have interesting coverage characteristics. It can be used as a geovisualisation tool to convey information to non-expert users about why the frequency of observation increases in higher latitudes or why observations of tropical regions have more cloud cover.

5. CHALLENGES AND OPEN ISSUES

Among the challenges encountered while developing the initial implementation of the Sentinel-2 Dashboard, the most demanding ones were issues regarding metadata harvesting from the Copernicus Open Access Hub, metadata quality and performance related issues.

Since the Sentinel-2 Dashboard operates its own data storage, the metadata database needs to be replicated from official Sentinel-2 data access portals; in this case the Copernicus Open Access Hub. Therefore, the database of the Sentinel-2 Dashboard is updated incrementally with a frequency of a few days. However, any scenes that are not published or are reprocessed are not considered in the analysis.

Similarly, there is no separate or additional validation of the metadata quality and correctness, since it would require too much processing overhead. Therefore, metadata information harvested from the Copernicus Open Access Hub are assumed to be complete and correct.

Regarding the analysis performance, the granule maps are the most demanding to implement because they are generated on the fly upon user request. At the beginning of each session, the Web application requests cached vector tiles of the granules. As soon as the user selects a topic (e.g. number of scenes, average cloud cover) the values are calculated by the database, grouped by granule id and joined with the granule footprint in the Web application. This guarantees that the most recent data is also taken into account. Even though the map is calculated on the fly based on the currently available 3.5 million metadata entries, the response time is usually less than five seconds, an acceptable performance for interactive usage scenarios. The database replicates the data for each granule into small individual tables. This allows calculating the temporal statistics for a granule without noticeable lag.

6. DISCUSSION AND CONCLUSION

The Sentinel-2 Dashboard generates statistics for all available Sentinel-2 scenes via intelligent metadata analysis and is a unique solution for aggregated image quality statistics. It allows users to use Sentinel-2 big data more intelligently by providing spatially explicit statistics about the availability and suitability of scenes. The results can be used as input for improving planning procedures, estimation of risks and interpreting results of analyses based on Sentinel-2 data. An example use case is a cloud-free mosaic or composite production that takes different revisit times between different regions into account.

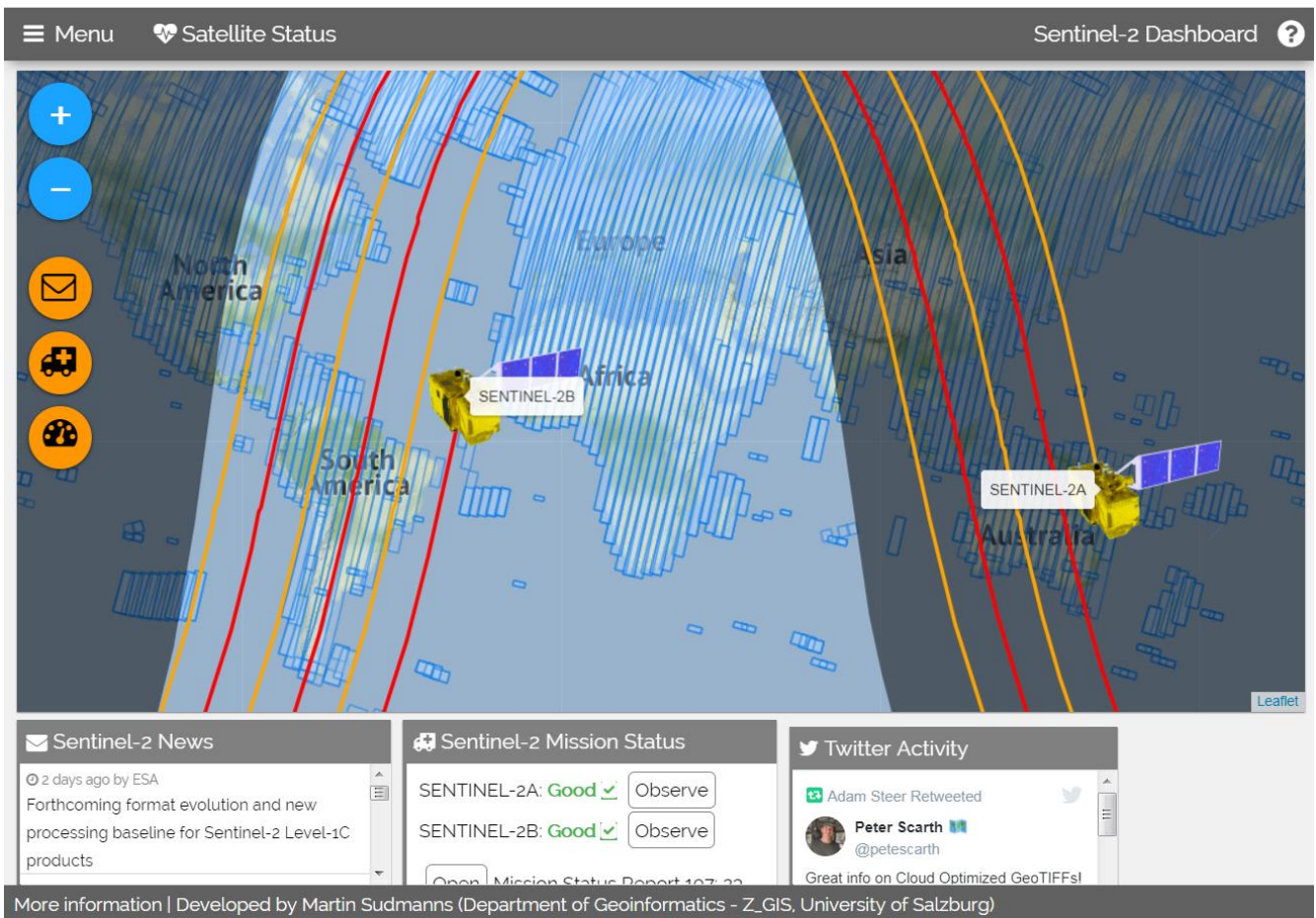


Figure 3: The Sentinel-2 Dashboard allows investigating current and predicted positions of the Sentinel-2 satellites.

In future work the Sentinel-2 Dashboard's functionality will be expanded. For example, additional data could be included: basemaps for land use / land cover; the current weather situation; a human settlement layer; or data about crisis/natural disasters. By using these additional data sets, the orbit paths with predicted cloud cover and geographic phenomena can be enriched, which can be translated into real-time information of interest to the public. For example, such information could be: "Sentinel-2A is now observing forest areas in Siberia, Russia", and "Sentinel-2B is currently collecting data about the living environment of 300 million people", "Sentinel-2A is collecting data about the flood that happened in Bangladesh". Estimated cloud cover of the weather forecasts can later be evaluated by the metadata of the Sentinel-2 images and vice versa. The orbits of other relevant satellites (e.g. Landsat) might be incorporated in the future for querying and predicting overlapping acquisitions in space and time. Moreover, the feasibility of re-using the software as a dashboard for Sentinel-1, Sentinel-3 or Landsat will be evaluated.

7. REFERENCES

- [1] ESA, *Sentinel-2. Mission Details*. <https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/sentinel-2>, 2017. Accessed 8 August 2017.
- [2] T. S. Kelso, *NORAD Two-Line Element Sets Current Data. CeresTrak Website*. <https://www.cesetrak.com/NORAD/elements/>, 2017 Accessed 12 October 2017.
- [3] J. Li and D. P. Roy, "A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring". *Remote Sensing* 9, 9, Basel, 2017.

OPTIMISING SENTINEL-2 IMAGE SELECTION IN A BIG DATA CONTEXT

P. Kempeneers and P. Soille

Joint Research Centre of the European Commission
via Fermi 2749, 21027 Ispra (VA), Italy

ABSTRACT

Processing large amounts of image data such as the Sentinel-2 archive is a computationally demanding task. However, for most applications, many of the images in the archive are redundant and do not contribute to the quality of the final result. An optimisation scheme is presented here that selects a subset of the Sentinel-2 archive in order to reduce the amount of processing, while retaining the quality of the resulting output. As a case study, we focused on the creation of a cloud free composite, covering the global land mass and based on the images acquired in 2016. The total amount of available images was 635,096 with an average of 34 overlapping images per tile. The selection of the optimal subset was based on quicklooks, which correspond to a spatial and spectral subset of the original Sentinel-2 products and are lossy compressed. They typically represent only 0.05% of the data archive volume. The result of the proposed selection scheme was a reduced set of images, with an average size of 2.55 images per tile.

Index Terms— image selection, Sentinel-2, Big Data, optimization

1. INTRODUCTION

The Copernicus program of the European Commission (EC) with its Sentinel satellites produces approximately 10 TB of Earth Observation (EO) data per day. This wealth of information, combined with a free full and open access policy provides new opportunities for applications in forestry, agriculture, and climate change monitoring, to name a few. An increasing number of platforms are being created that address the storage and processing of these data, both from the institutional and the private sector. The need for computing power that can process the amount of available data can only be expected to increase. The EC has launched an initiative earlier this year to develop Copernicus Data and Information Access Services (DIAS) that facilitate access to these data and allow for a scalable computing environment.

Nevertheless, computing power comes at a cost, both financially and environmentally [7]. A typical data center with thousands of servers may consume as much energy as 25,000 households [1]. Processing large amounts of data also takes valuable time, not only from the servers, but also from the

users that have to wait for their results. Our contribution with this paper is to provide means to reduce the computations in a twofold approach. First, we minimise the number of images that must be processed without impacting the quality of the final result. Second, instead of dealing with the original images at full resolution, the selection criterion is based on a sample only. In our case, this was performed by considering the quicklooks of the images, that is a spatial and spectral subset of the original images.

The reduction of the number of images is driven by the fact that for most applications, not all acquired images are equally valid. Often, the useful information can be obtained from a reduced selection of the images. The challenge that was tackled in this paper, was to find the minimum set of images that maximise the information content for the problem at hand. A fitness criterion was therefore defined that evaluates each image. The underlying idea is that the computational cost of the optimisation function is considerably less than the processing algorithm that is needed to produce the application results.

2. MATERIALS AND INFRASTRUCTURE

As an illustration of the optimised image selection in a Big Data context, we focused on the application of the creation of a global cloud free satellite image composite of the land surface. This application can serve other derived products, for instance the identification of human settlements [3]. The satellite image composite was based on optical remote sensing data acquired with the Sentinel-2 sensors. Sentinel-2A and Sentinel-2B are a constellation of two identical optical sensors with 13 spectral bands. Together, they almost cover the entire land surface of the globe every five days. The original images have a spatial resolution of 10 to 60 m, depending on the spectral band. They daily generate of about 1.6 TB of compressed raw image data [2]. The European Space Agency (ESA) provides the Sentinel-2 images in 100 × 100 km tiles according to the Military Grid Reference System (MGRS). In addition to the original images, ESA provides true colour quicklooks at a reduced spatial resolution of 320 m (see Fig. 1a).

The infrastructure that is available in house for this analysis is based on commodity hardware and open source soft-

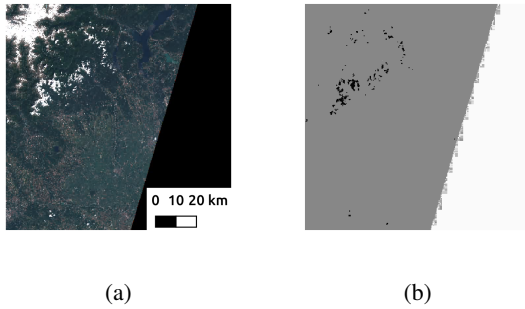


Fig. 1: Example of quicklook (a) and mask defining cloud and data domain (c)

ware [5, 6]. The operating system is based on Linux that is run on both the the storage and processing nodes. There are 39 processing nodes with 968 cores and 15.3 TB of RAM in total. The full storage, currently at 1.9 TB, is served through a distributed file system (DFS). The implementation of the DFS is EOS, which has been developed by the European Organisation for Nuclear Research (CERN). For the workload manager, HTCondor was selected. It is particularly fit for the high throughput computing we are typically faced with in Earth observation data processing.

3. METHODS

Due to cloud cover that occurs in different parts of the globe, multiple acquisitions acquired at different dates are required for a global cloud free composite. For some regions, images acquired at a single acquisition date in cloud free conditions can be found. For other regions, due to persistent cloud cover, overlapping images acquired at different times must be combined in order to obtain a cloud free composite. For instance, in equatorial South America, the Congo River basin in Africa, and Southeast Asia, the probability of cloudy observations is typically more than 80% [8].

The 290 km wide swath of any given relative orbit is only partially covering the MGRS tiles intersecting the boundary of the swath. We therefore created a binary mask that corresponds to the data domain, i.e., pixels where valid data are present. The mask was then combined with a rasterized version of the cloud mask, which was provided by ESA in vector format (see Fig. 1b).

The optimisation of selecting the minimum set of images that covers the data domain is schematically illustrated in Fig. 3. The example shows how we can fall into a local minimum that results in a sub-optimal set. There are three overlapping images 1-3 for which cloud masks have been created (mask1-mask3 respectively, see Fig. 2). Starting from the least cloudy image 1, three images (1, 2, and 3) are required to obtain a cloud free composite. However, the minimum set is composed of two images (2 and 3). A

global optimal set can be selected using an exhaustive search through all the combination of all N available images. However, the number of combinations grows exponentially with the number of available images (2^N). We used the sequential floating forward search algorithm (SFFS) [4], which avoids the local minimum and is able to find the correct set of images (2 and 3).

Algorithm 1 Sequential Floating Forward Search

Inputs: overlapping image set $Y = \{y_j | j = 1, \dots, N\}$ and maximum number of selected images M_{\max} , where $1 \leq M_{\max} \leq N$

Output: selected images $X = \{x_j | j = 1, \dots, M\}$, $x_j \in Y$ and $1 \leq M \leq M_{\max}$

Initialization:

$X_0 = \emptyset$; $k = 0$

repeat

Inclusion {Add the most significant image with respect to X_k }

$x_j^+ = \arg \max_{x_j \in Y \setminus X_k} J(X_k \cup x_j)$

$X_{k+1} = X_k \cup x_j^+$; $k = k + 1$

repeat

Conditional exclusion {Try to remove images}

$x_j^- = \arg \max_{x_j \in X_k} J(X_k \setminus x_j)$ {the least significant image in X_k }

$X_{k-1} = X_k \setminus x_j^-$; $k = k - 1$

until $J(X_k \setminus x_j^-) < J(X_{k-1})$

until X_k is cloud free or $k = M_{\max}$

$X = X_k$

The iterative algorithm is listed in Algorithm 1. It selects the minimum set of M images (X_M) from the entire set of overlapping images Y for a specific tile to obtain a cloud free composite. We start with an empty set X_0 . At each iteration, we try to include the most significant image with respect to X_k , with k the current number of images in the selected set. The objective function $J(X_k)$ is based on the number of cloud free pixels within the data domain and their mean value in the blue band. To avoid a local minimum, we try to remove the least significant image in X_k . If the objective function of the reduced set is lower than that of the set of the previous iteration (X_{k-1}), we replace the current set with the reduced set. In the original algorithm proposed [4], the iteration continues until the the set X_k reaches the required dimension. Here, we slightly adapted the algorithm and stop the iteration when a cloud free composite is obtained. In case a single cloud free image covers the entire tile, the number of images is one ($M = 1$). Notice that some pixels that cover highly reflective areas on the ground can result in consistent commission errors of the cloud detection algorithm. We therefore constrained the dimension of the set to a maximum value ($k \leq M_{\max}$). Without this constraint, the algorithm would select the entire overlapping image set ($M = N$), in an attempt to obtain a cloud free composite.

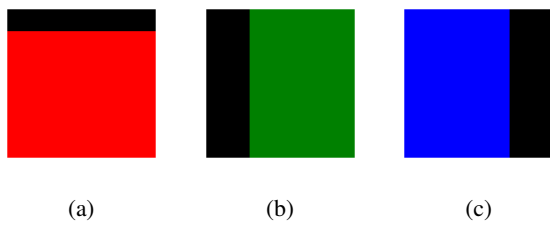


Fig. 2: Cloud masks (colored area is within data domain and cloud free) for three overlapping images: mask1 (a: 15% no data), mask2 (b: 29% no data) and mask3 (c: 29% no data).

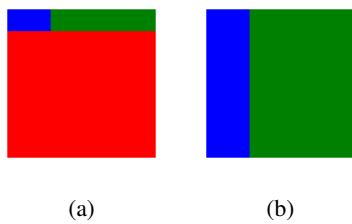


Fig. 3: Image selection for cloud free composite within data domain using sub-optimal set of three images (a) and optimal set of two images (b). Colors represent the selected image from Fig. 2.

4. RESULTS

The optimal list of images was selected for the 30,807 distinct MGRS tiles, corresponding to 635,096 S2 image tiles. The total processing time on the JEODPP [5] cluster was 20 hours using 968 cores. The blue area in Fig. 4 represents the CPU user time as a percentage of the full cluster capacity.

Another interesting outcome of this work was the insight in the number of overlapping tiles required (see Table 1). For 29% of all tiles, a single image was selected ($M = 1$). The remaining tiles required a combination of overlapping images to obtain a cloud free composite. As much as 29% of the

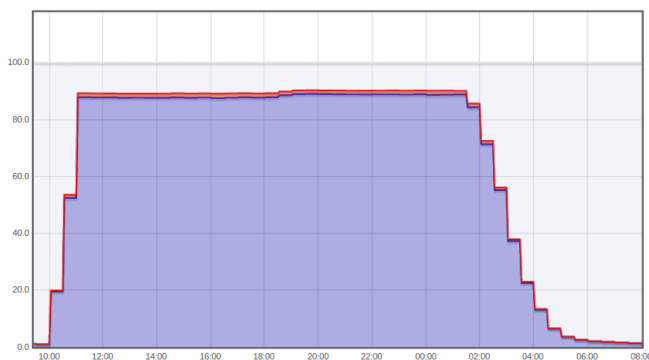


Fig. 4: Processing time spent for the global image selection using 968 cores.

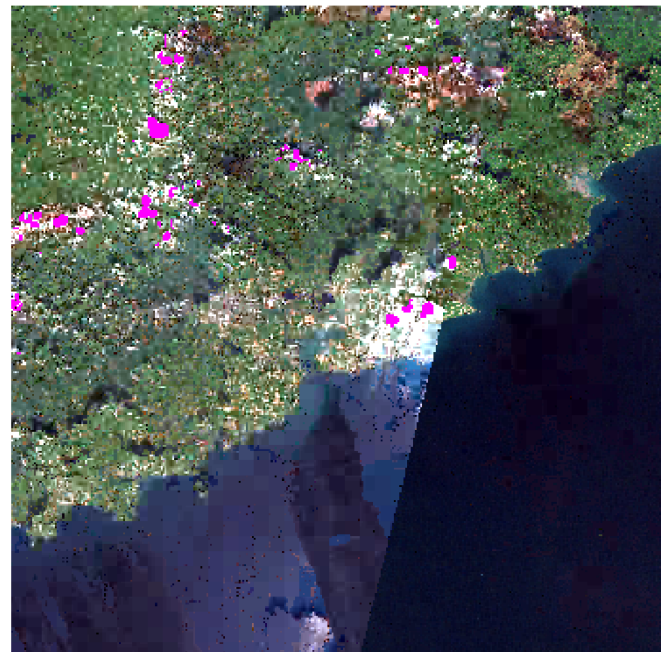


Fig. 5: Minimum composite of five overlapping images for the tile T29UNT, south of Ireland (city of Cork), with persistent cloud cover (pixels in pink).

Table 1: Distribution in percentage of number of overlapping images (M) required to obtain a cloud free composite.

M	1	2	3	4	5 or more
percentage	29%	27%	11%	5%	29%

tiles remained cloudy to some extent even after selecting five overlapping scenes. As example, in Fig. 5 a composite of five overlapping images is shown for a tile in Cork, south of Ireland. The world composite (see Fig. 6), shows that most tiles were effectively cloud free.

5. CONCLUSION

An optimisation scheme was presented that selects a subset of the Sentinel-2 archive in order to reduce the amount of processing, while retaining the quality of the resulting output. As a case study, we focused on the selection of a cloud free composite, covering the global land mass and based on the images acquired in 2016. The selection of the optimal subset was based on quicklooks, which correspond to a spatial and spectral subset of the original Sentinel-2 products and are lossy compressed. They typically represent only 0.05% of the data archive volume. The result of the proposed selection scheme was a reduced set of overlapping images, with an average size of 2.55 images per tile. Based on the minimum set, the cloud free global composite could then be produced at full spatial and spectral resolution. Using the selected set of images instead

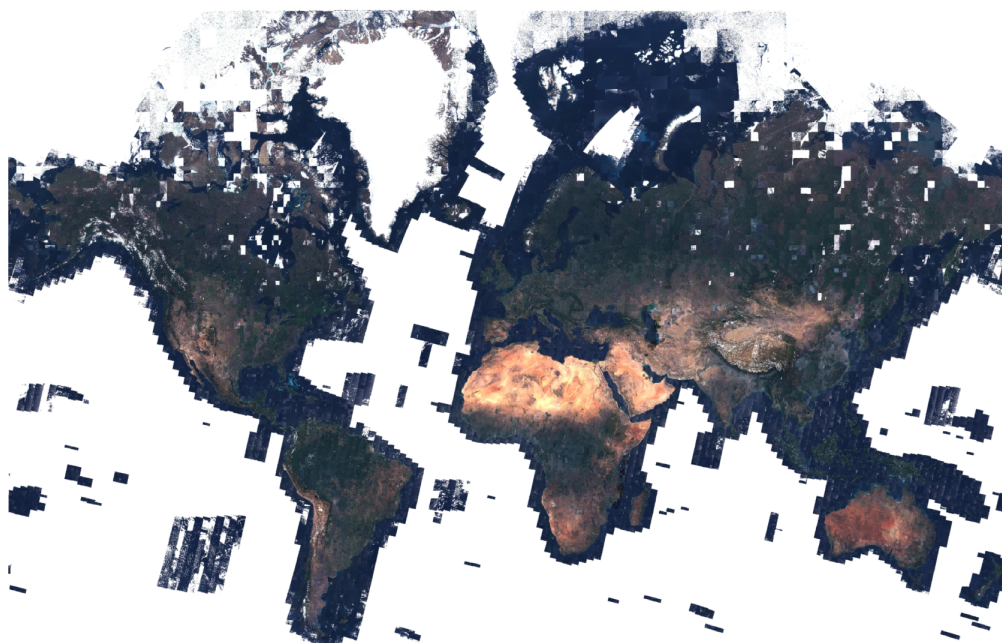


Fig. 6: World composite based on selected Sentinel quicklooks.

of the full set, only 7.5% of the images had to be processed in their full spatial and spectral resolution.

6. REFERENCES

- [1] Dayarathna, M. et al. “Data center energy consumption modeling: A survey”. *IEEE Communications Surveys & Tutorials* 18.1 (2016), pp. 732–794. DOI: 10.1109/COMST.2015.2481183.
- [2] Drusch, M. et al. “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services”. *Remote Sensing of Environment* 120 (2012), pp. 25–36. DOI: 10.1016/j.rse.2011.11.026.
- [3] Pesaresi, M. et al. “Assessment of the added-value of sentinel-2 for detecting built-up areas”. *Remote Sensing* 8.4 (2016), p. 299. DOI: 10.3390/rs8040299.
- [4] Pudil, P. et al. “Floating search methods in feature selection”. *Pattern recognition letters* 15.11 (1994), pp. 1119–1125. DOI: 10.1016/S0167-8655(99)00083-5.
- [5] Soille, P. et al. “The JRC earth observation data and processing platform”. In: *Proceedings of the Conference on Big Data from Space (BiDS’17), Toulouse, France*. 2017.
- [6] Soille, P. et al. “Towards a JRC earth observation data and processing platform”. In: *Proceedings of the Conference on Big Data from Space (BiDS’16), Santa Cruz de Tenerife, Spain*. 2016, pp. 15–17. DOI: 10.2788/854791.
- [7] Whitehead, B. et al. “Assessing the environmental impact of data centres part 1: Background, energy use and metrics”. *Building and Environment* 82 (2014), pp. 151–159. DOI: 10.1016/j.buildenv.2014.08.021.
- [8] Wilson, A. M. and Jetz, W. “Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions”. *PLoS biology* 14.3 (2016), e1002415. DOI: 10.1371/journal.pbio.1002415.

SCALABLE CLOUD-BASED COMPUTATION OF CONSISTENT SURFACE REFLECTANCE MOSAICS AT 10M FROM SENTINEL-2 AND LANDSAT-8 MISSIONS

Konstantinos Karantzalos and Athanasios Karmas

Remote Sensing Laboratory
National Technical University of Athens
Zographou campus, 15780, Athens, Greece

karank@central.ntua.gr thanasis.karmas@gmail.com

ABSTRACT

Emerging critical monitoring applications in land use, land management, agriculture, security as well as in other sectors require frequent high resolution earth observation data. Leveraging the time domain at high resolution scales could, under specific conditions, form a breakthrough in Copernicus EO data exploitation and thus create solid foundation for novel products and services. To this end, in this paper we present a scalable approach for the computation of consistent surface reflectance surfaces at 10m spatial resolution based on EO data streams from Sentinel-2 and Landsat-8 missions. An automated, modular, distributed workflow has been designed and developed based on state-of-the-art computer vision algorithms. These algorithms are able to harmonize the surface reflectance data, to co-register and resample all datasets (over the Sentinel-2 tiling system) as well as to efficiently pan-sharpen lower resolution spectral bands and imagery at 10m for the final production of multitemporal mosaics. Initial implementations exploit inter-CPU parallelism along with the underlying infrastructure's distributed computation environment (computer cluster). For further optimizations GPU implementations are highly considered.

Index Terms— Co-registration, Pansharpening, Surface reflectance, Mosaics, Distributed processing

1. INTRODUCTION

Apart from the spatial and spectral resolution, the temporal one is crucial for numerous geospatial applications. In particular, several applications in land use, land management, agriculture, security as well as in other sectors require frequent (e.g., weekly) high resolution earth observation data. Through advanced analytics and machine learning at the image domain or based on time series analysis, changes can be detected, classified and recognized. However, in order to achieve this, the main prerequisite is the frequent and consistent data availability. In particular, from the currently feasible inter-annual analysis, disturbance, compositional change and land use the Earth Observation community can move a step

further towards intra-annual analytics, modelling, land cover change, crop phenology, vegetation condition, compositing and many more applications. Leveraging the time domain at high resolution scales (e.g., at 10 meters) could under specific conditions form a breakthrough in Copernicus EO data exploitation and thus create solid foundation for novel products and services.

Currently, one way to leverage the time domain with high resolution datasets is through the combination of the Copernicus Sentinel-2, the Landsat-8 and the Copernicus Contributing Missions (Mission Group #2). In particular, by merging Sentinel-2 (S2) and Landsat-8 (L8) data streams, a coverage cycle of less than 5-days can be achieved in most geographical regions [1]. Such a dataset would enable the delivery of around 50 images in a six month period, while by assuming that around 40% will be covered with clouds (e.g., Conterminous United States [2]) it would enable the delivery of time series with around 30 cloud-free observations. Towards achieving this, consistent, harmonized surface reflectance datasets must be systematically produced in order to generate seamless multitemporal, multispectral mosaics at 10m resolution. Consistency is required both in the spatial and spectral domain by addressing all radiometric and geometric challenges. However, at the present time and despite all research efforts, S2 and L8 datasets are not geometrically aligned (up to 40m spatial displacements exist). Moreover, misalignment cases between S2 tiles occur and not-calibrated surface reflectance data exist [3].

With the aim to answer all these crucial challenges, a comprehensive and scalable framework by the name Earth10 is established in the present work. The framework is capable of providing radiometric and geometric consistent surface reflectance mosaics at 10m by merging and cross-calibrating S2, L8 as well as other datasets from Copernicus Contributing Missions. An automated processing pipeline has been designed and developed based on state-of-the-art computer vision algorithms that forms the core of the established framework. Earth10 is able to harmonize the surface reflectance data, to co-register and resample all datasets (over the S2

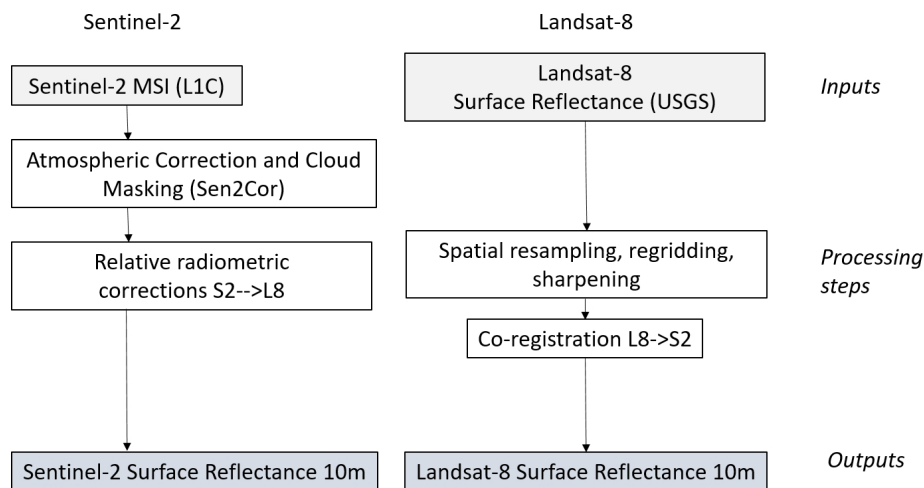


Fig. 1. A flowchart of the developed methodology for computing the Earth10 products

tiling system) as well as to efficiently sharpen lower resolution data at 10m and finally produce multitemporal mosaics.

2. RELATED WORK

Current research and development efforts have been concentrated in the inter-comparison of surface reflectance (*e.g.* ACIX exercise) as well as the development of a harmonized surface reflectance product from University of Maryland (UMD) and NASA¹. However, the latter is focusing on a Landsat-like product at 30m spatial resolution and therefore all the significant spatial information derived from S2 is lost. Furthermore, current analytics on both S2 and L8 datasets for time series analysis or change detection are based neither on consistent radiometry nor geometry and as a consequence results are problematic and inconsistent, thus leading to significant signal variation and false alarms ([4], [5]). Moreover, other Copernicus contributing missions are not considered.

Earth10 provides consistent products at 10m spatial resolution by exploiting cutting-edge pan-sharpening and data fusion algorithms. The framework addresses efficiently the mis-registration issues and as a consequence is able to deliver consistent spatio-temporal surface reflectance surfaces.

3. METHODOLOGY

Earth10's surface reflectance mosaics at 10m are based on S2, L8 and Copernicus Contributing Missions, (mainly but not limited to) Mission Group 2 - Optical HR1/2. Currently, for the L8 data streams, Earth10 is based on the standard surface reflectance product that is delivered from the LaSRC (Land Surface Reflectance Code) algorithm USGS² (Figure 1). As far as the S2 data streams are concerned, Earth10 utilizes the standard Level-1C product delivered from ESA

and uses the Sen2Cor³ processor for the calculation of the necessary surface reflectance data. Software modules implemented in *Python* are responsible for the automated downloading, archiving, cataloguing and transformation of the imagery datasets.

The aforementioned surface reflectance products are the main requirement. However, there are software modules that can also perform atmospheric corrections at the raw satellite datasets in order to efficiently address simultaneously relative radiometric corrections and cross-calibration. The current effort is to integrate these software modules to the Earth10's workflow. Furthermore, reflectance in-situ datasets from global calibration and benchmark initiatives can be also integrated in order to constantly validate the generated mosaics.

Earth10 is closely following the outcomes of the Atmospheric Correction Inter-comparison Exercise (ACIX) as well as the harmonized surface reflectance product from UMD and NASA. At the moment, concentrating more on the performance over the land and not water regions the Sentinel-2 surface reflectance outputs are slightly adjusted towards matching the relative reflectance of the atmospherically corrected L8. In Figure 1, certain software modules that orchestrate and perform all processing tasks like geometric and radiometric corrections are presented. This is the same for the case of other data from other contributing missions (*e.g.* Deimos-1, Pleiades).

Regarding all geometric issues ([3], [6]), state-of-the-art automated registration algorithms are addressing all misalignment cases along with the simultaneous resampling of all the involved surface reflectance datasets upon the S2 Tiling Grid Reference System. The spatial resampling process has been designed and developed specifically for the targeted surface reflectance product at 10m. The process is based on an area weighted average strategy when increasing the resolution of

¹<https://hls.gsfc.nasa.gov/>

²<https://www.usgs.gov/>

³<http://step.esa.int/main/third-party-plugins-2/sen2cor/>

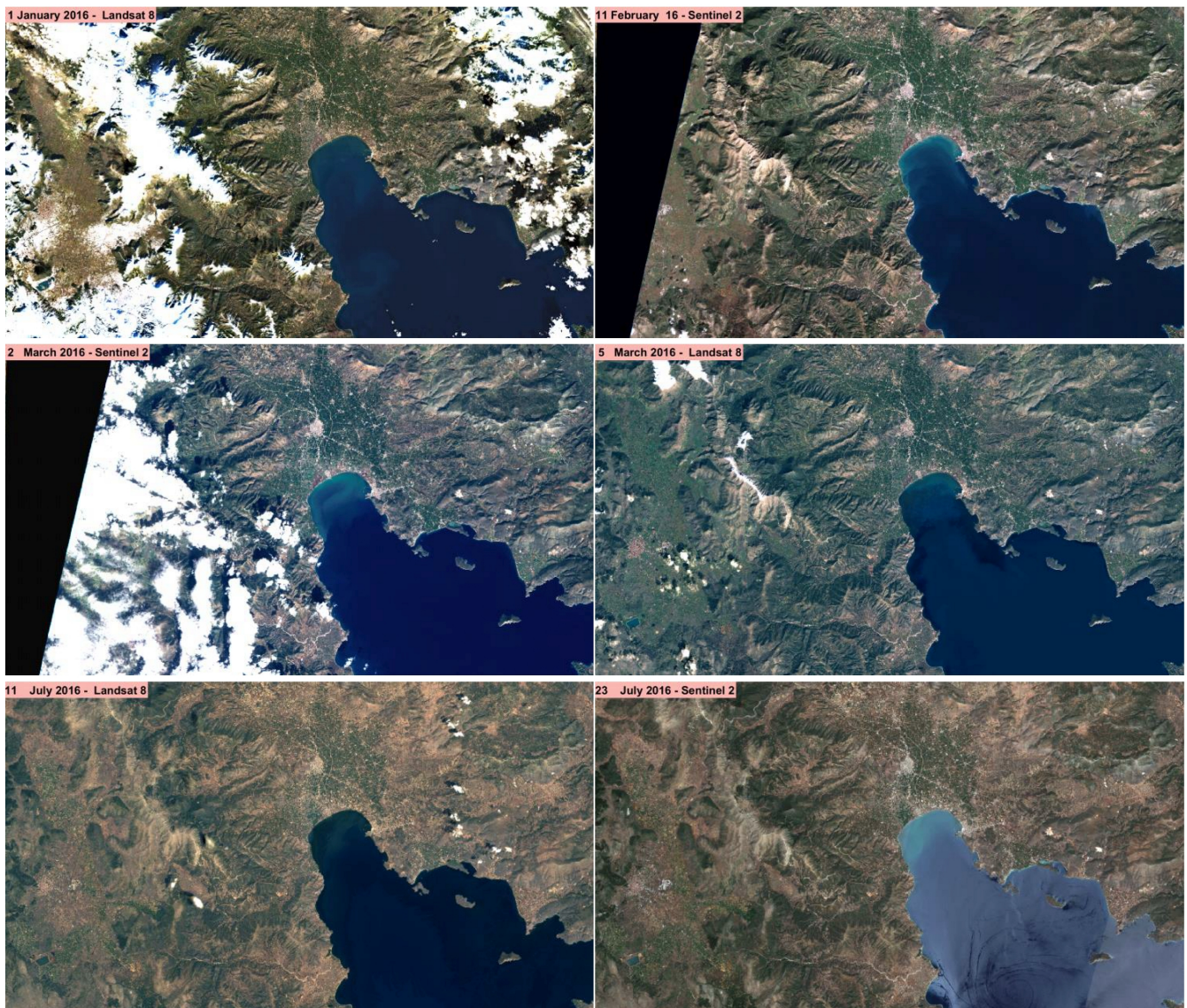


Fig. 2. Example of co-registered, pansharpened Sentinel-2 and Landsat-8 data at 10m spatial resolution. More experimental results here: <http://users.ntua.gr/karank/Earth10.html>

L8 data streams to the S2 dataset resolution (from 30m to 10m). The spatial resampling process has been implemented in *Python* and exploits a computer system's multiprocessing capabilities for maximum efficiency.

Apart from the four S2 spectral bands at 10m (R,G,B,NIR) all the other spectral bands from S2 as well as L8 are consistently pansharpened at 10 meter resolution ([7], [8]).

Earth10 is forced to address current big data challenges and thus all critical components of the framework are modular and fully scalable based on automated processing pipelines implemented on the ARIS⁴ supercomputer infrastructure. ARIS is the name of the Greek supercomputer, deployed and operated by GRNET(Greek Research and Technology Net-

work) in Athens. ARIS consists of 532 computational nodes⁵. As a consequence, the developed processing pipelines can be easily migrated on various third party cloud computing platforms. Automated pipelines are built from various architecture-agnostic components (*i.e.* virtualized components) that are able to handle all the data transferring, management, archiving, (pre-)processing demands through the exploitation of both well known on-demand cloud computing platforms as well as in-house high performance computing solutions. As most of the algorithms involved are inherently parallel, initial implementations exploit inter-CPU parallelism as well base infrastructure's distributed computing environment (*i.e.* computer cluster). Earth10 is under inten-

⁴<https://hpc.grnet.gr/en/>

⁵<http://doc.aris.grnet.gr/hardware/>

sive validation, while optimizations are an ongoing process primary through GPU implementations. The modularity of the developed software ensures the potential for integration with distributed geospatial processing systems that can guarantee the near real-time delivery of results when relative small regions are concerned due to the optimized computer vision algorithms (relative low CPU, GPU core hours per tile) and optimized data access functions.

4. EXPERIMENTAL RESULTS AND EVALUATION

Experimental results are presented in Figure 2, as well as in a demonstration web page⁶ with animations for showcasing the merged multitemporal imagery at 10 meter spatial resolution. In particular, in Figure 2 an example of Earth10's surface reflectance product at 10m spatial resolution is displayed. One can observe side by side S2 surface reflectance imagery along with L8 pansharpened and harmonized imagery as produced by the Earth10's workflow, co-registered with each other so that no spatial displacements between the two different datasets occur.

More specifically, one can observe the similar intensity values of the final outputs *i.e.*, L8 and S2 reflectance at 10m. This is the case independently of the seasonal variability as it can be observed in Figure 2 top (winter snow cover), Figure 2 middle (spring serious cloud cover) and Figure 2 bottom (summer clear acquisition).

Moreover, the intensity values of the final product are extremely harmonized not only when land areas are concerned but also when water filled areas are the case (*e.g.* sea in Figure 2). It is evident that the L8 data streams are filled with significant spatial information (from 30m to 10m / pixel) that can prove to be critical for several land, urban, agriculture, *etc.* monitoring applications. The final product is also suitable for high resolution (10m), high frequency time series analysis as it is able to exploit both datasets and amend for expected cloud coverage that may cancel a significant percentage of satellite acquisitions.

5. CONCLUSION

In this paper a scalable approach for the computation of consistent surface reflectance surfaces at 10m spatial resolution was presented. The product is currently based on EO data streams from S2 and L8 missions. An automated, modular and distributed workflow has been designed and developed based on state-of-the-art computer vision algorithms. Through the workflow the datasets are subjected into several transformations in order to solve issues of harmonizing the various surface reflectance data, of co-registering and resampling all datasets (over the S2 tiling system), of efficiently pansharpening lower resolution spectral bands and imagery at

10m for the final production of multitemporal mosaics. Initial implementations exploit inter-CPU parallelism as well as base infrastructure's parallelism through computer clusters, while further optimizations are considered through GPU implementations.

6. REFERENCES

- [1] Jian Li and David P. Roy, "A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring," *Remote Sens.*, vol. 9, no. 9, 2017.
- [2] Kovalskyy V and David P. Roy, "A one year Landsat 8 conterminous United States study of cirrus and non-cirrus clouds," *Remote Sens.*, vol. 7, 2015.
- [3] James Storey, David P. Roy, Jeffrey Masek, Ferran Gascon, John Dwyer, and Michael Choate, "A note on the temporary misregistration of Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi Spectral Instrument (MSI) imagery," *Remote Sensing of Environment*, vol. 186, pp. 121 – 122, 2016.
- [4] Lin Yan, David P. Roy, Hankui Zhang, Jian Li, and Haiyan Huang, "An Automated Approach for Sub-Pixel Registration of Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi Spectral Instrument (MSI) Imagery," *Remote Sensing*, vol. 8, no. 6, 2016.
- [5] Sergii Skakun, Jean-Claude Roger, Eric F. Vermote, Jeffrey G. Masek, and Christopher O. Justice, "Automatic sub-pixel co-registration of Landsat-8 Operational Land Imager and Sentinel-2A Multi-Spectral Instrument images using phase correlation and machine learning based mapping," *International Journal of Digital Earth*, vol. 0, no. 0, pp. 1–17, 2017.
- [6] Christos Platias, Maria Vakalopoulou, and Konstantinos Karantzaos, "Fully Automated Sentinel-2 Data Registration to Various Multi-sensor Earth Observation Imaging Datasets," in *European Space Agency (ESA), Living Planet Symposium*, 2016.
- [7] Aristeidis Vaiopoulos and Konstantinos Karantzaos, "Pansharpening on the Narrow VNIR and SWIR Spectral Bands of Sentinel-2," in *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016.
- [8] Hankui K. Zhang and David P. Roy, "Computationally Inexpensive Landsat 8 Operational Land Imager (OLI) Pansharpening," *Remote Sens.*, vol. 8, no. 3, 2016.

⁶<http://users.ntua.gr/karank/Earth10.html>

A MODEL-DRIVEN BIG DATA ARCHITECTURE FOR PLANETARY DATA ARCHIVES AND RESEARCH

Daniel J. Crichton¹, J. Steven Hughes¹, Sean Hardman¹, Emily Law¹, Thomas C. Stein², Reta Beebe³

¹Jet Propulsion Laboratory, ²Washington University in St. Louis, ³New Mexico State University

ABSTRACT

The NASA Planetary Data System captures, archives, and distributes data from robotic exploration of the solar system. In supporting this mission, it has developed an innovative architectural approach called “PDS4” to support the highly diverse set of heterogeneous data from more than 600 instruments. The PDS is implemented as a set of distributed archives with different “nodes” managing repositories for this federated system. To enable the federated approach, the PDS uses an information model to drive configuration of its archive and services, enabling it to evolve as data from the mission evolves. This approach has also enabled the PDS to work with and share its standards and architectures with the international community through the International Planetary Data Alliance.

Index Terms— Planetary science, data archiving, interoperability, PDS, data modeling

1. INTRODUCTION

The NASA Planetary Data System (PDS) is NASA’s official archive for capturing, managing and distributing observational science data from robotic exploration of the solar system [1]. It is organized around science disciplines called “nodes”, with each node managing portions of the peta-scale archive. The Engineering Node, located at the Jet Propulsion Laboratory, provides overall architecture, development and coordination of standards, software, and operations. Six science discipline nodes are managed by leading planetary scientists who each provide expertise in working with their communities to ensure the PDS can support the scientific needs of the missions and users for their disciplines.

In 2010, the PDS began the largest standards and software upgrade in its history called “PDS4” [2]. PDS4 was architected with core principles, applying years of experience and lessons learned working with scientific data returned from robotic solar system missions. In addition to applying those lessons learned, the PDS development team was able to take advantage of modern software and data architecture approaches and emerging information technologies which have enabled the capture, management, discovery, and distribution of data from planetary science archives worldwide. What has emerged is a foundational set of standards,

services, and common tools to construct and enable interoperability of planetary science archives from distributed repositories.

At the heart of the architecture is an information model captured and managed using modern ontology modeling tools. The information model describes the planetary data architecture including the data itself, structure, and organization providing standard structures, dictionaries, and relationships. Given the diversity and distribution of planetary science data, it is critical that the PDS uses this information model driven approach to configure and drive consistency in its data management, search, and analytic capabilities as shown in figure 1. The information model is managed by the PDS Engineering Node to ensure it evolves as a common semantic basis for integration of data across the highly distributed PDS.

2. TECHICAL APPROACH

The PDS4 reference architecture is decomposed into three parts:

- Information Model – provides the definition of data and their relationships
- Software Services – provides the system services to manage, search, and distribute data
- Tools – supports validation and use of the data

Science data including metadata, data, and ancillary information generated, captured, and managed by planetary missions and archived in the PDS are defined and constrained by the information model. The information model includes classes, attributes, and relationships that define the metadata used to describe the data captured from a planetary observation. The information model includes a multi-level governance approach (see figure 2) which requires a set of metadata at the common level but can be extended to support different planetary science disciplines (imaging, geosciences, rings, fields and particles, small bodies, etc.) as well as mission uses following an object-oriented modeling approach. While the common and extensions can each represent their own local models, it’s the integration of all of these local(?) models, including international versions, which constitute the broad planetary data architecture under PDS4.

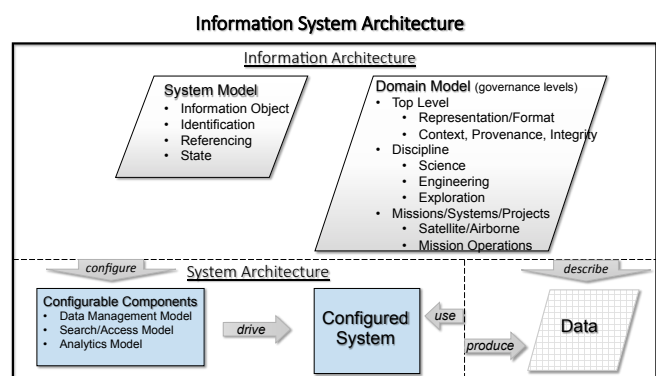


FIGURE 1: MODEL-DRIVEN ARCHITECTURE

The information model is explicitly captured as an ontology using the Protégé modeling tool software. While the multi-level governance approach allows for extension by different stewards, they still must adhere to a set of standards and conventions to ensure there is one overarching model to describe the domain. This is critical to ultimately drive management, discovery, and analytics. The standards ensure that each local model inherit the properties of the common PDS4 information model including data dictionaries and more generalized classes. The data dictionaries are organized according to the ISO/IEC 11179 standard. Generalized classes describe platforms, observations, data, and other types of information captured in the archives. Extensions to these allow for the specialization. For example, PDS4 provides a common class to capture arrays which represent a significant portion of scientific data returned from instruments. These can be extended to support specialized arrays that represent different types of imaging formats (e.g., multi-dimensional color images.) Similarly, new classes that represent specialized metadata to further describe a planetary observation can be specified to enhance discovery and use of the data. Such examples include cartography information that would enhance imaging and other types of spatial data for different planetary targets (e.g., Mars, Vesta, etc.). All PDS metadata are captured as XML representations of the science data.

Decades of experience managing planetary data archives have led the PDS to recognize that software should be driven by this model rather than be loosely coupled. This is important due to the continued increasing diversity of data types within the PDS. The information model produces an explicit set of artifacts that can be used to configure software (Figure 3) such as XML schemas and Schematron (reference) to support syntactic and semantic validation of the XML metadata. Furthermore, the artifacts include RDF, JSON, and other outputs that can be used in software to describe data classes and their attributes. Software services including

harvest, registration, search, transformation, and distribution support the archive lifecycle functions. These are all configured by the output of the model so that no local customization is required when extensions or improvements to the model are made.

Given the federated PDS architecture and diversity of data, a distributed registry and search approach is used to allow different nodes to manage portions of the archive. The PDS4 XML-based metadata are extracted and registered within the registry services based on the PDS4 information model. This allows tracking of data in the archive. Data are further extracted and indexed by an open source search engine implementation using Apache Solr (reference) to form the PDS4 search service, enabling free text search using terms from the PDS4 information model. This registry search approach allows registration and search at international scales. Today, for example, NASA and the European Space Agency (ESA) already are exposing their metadata to this search engine. As part of distribution of the data, the PDS has developed transformation capabilities to convert PDS archival data into other formats.

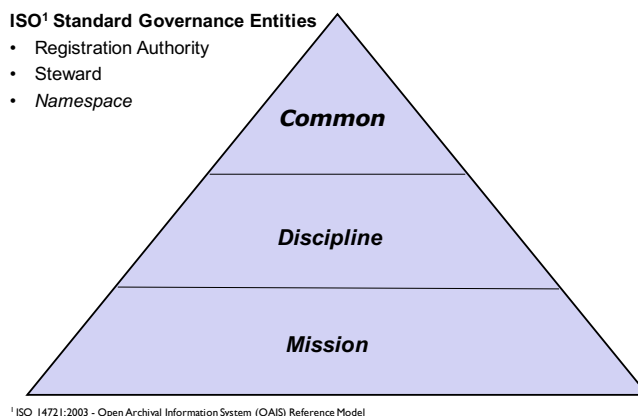


FIGURE 2: MULTI-LEVEL GOVERNANCE

PDS4 also provides a set of tools that are configured by the information model. A principal tool is the *validate tool*. The validate tool provides semantic, syntactic, content, and referential integrity validation. The tool uses the information model to configure itself to validate the metadata and key aspects of the data itself (e.g., validating table structures, etc.). Other tools and libraries support designing the metadata labels as well as reading, inspecting, and using specific types of data.

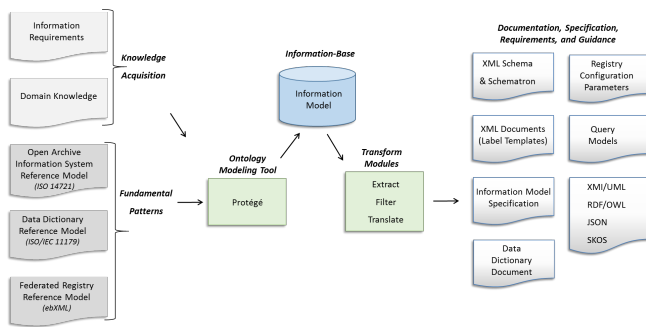


FIGURE 3: MODEL-DRIVEN PROCESS

3. APPLICATIONS

Early in the PDS4 development, the PDS selected two NASA missions as drivers to be used to validate the PDS4 approach: LADEE and MAVEN. These missions provide a set of data products that were used to validate the model-driven architecture approach without overwhelming the development teams. Using two different sets was important to ensure that similarities and differences were understood and addressed. The teams, working with PDS nodes, designed data products starting with V1.0 which was released at the end of 2013. Developing these in concert with the model allowed for significant learning and evolution paved the way for adoption for others.

The PDS partnered with international agencies through the International Planetary Data Alliance (IPDA) to coordinate the architecture, design, and implementation to ensure that PDS4 is architected as a world-wide standard and platform for archive development and interoperability. The IPDA, an international consortium established in 2006 for the development of compatible planetary science archives, has representatives from space agencies around the world working to ensure that planetary data standards can be successfully implemented and used by their scientific communities. The adoption of PDS4 lays a framework for interoperability going forward.

Given the evolving requirements and diverse community of both archives and users, an agile software development methodology known as the “Evolutionary Software Development Lifecycle” (reference) was chosen. This led to incremental releases of increasing capability which were matched against emerging mission and user needs. To date, the PDS has now performed 9 operational releases of PDS4 with adoption or planned adoption by 14 missions world-wide (Table 1). PDS data holdings during PDS4 development increased from approximately 200 TBs in 2010 to approximately 1.3 PBs of data today, bringing it into the era of big data.

TABLE 1: MISSIONS USING PDS4

Mission	Agency	Launch Date
---------	--------	-------------

LADEE	NASA	2013
MAVEN	NASA	2013
Hayabusa 2	JAXA	2014
OSIRIS-REx	NASA	2016
ExoMars TGO	ESA	2016
InSight	NASA	2018
Chandrayaan-2	ISRO	2018
BepiColombo	ESA/JAXA	2018
ExoMars Rover Surface Platform	ESA	2020
Mars 2020	NASA	2020
LUCY	NASA	2021
JUICE	ESA	2022
Psyche	NASA	2022
Europa Clipper	NASA	2020s

4. FUTURE

The development of PDS4 has not only focused on the construction of compatible archives, but also on increasing access and use of the data in the big data era. This directly responds to the recommendations of the Planetary Science Decadal Survey (2013-2022): “to support the ongoing effort to evolve the Planetary Data System to an effective online resource for the NASA and international communities” [3]. The foundation laid by the PDS4 standards, software services, and tools positions the PDS to develop and adopt new approaches and technologies to enable users to effectively search, extract, integrate, and analyze with the wealth of observational data across international boundaries.

5. CONCLUSION

PDS4 has provided several innovations for NASA in the development of distributed archives and data systems. In particular, it pioneered the development on an information-model driven approach to handle the diversity of the planetary science discipline through PDS4. This has proven to be both extensible and scalable as NASA and the international community have begun applications to active missions. The architecture approach is enabling PDS and the international community to support the diversity of the data across disciplines and missions while driving towards a research data platform for planetary science. Looking forward, as the PDS evolves to enable greater access and use of its data, we anticipate PDS4 providing an excellent foundation to build out this research platform including new user access services, increased searching at international scale, and providing improved approaches for users to discover and use data not only in the PDS, but across world-wide planetary archives.

6. REFERENCES

[1] *Special Issue: The Planetary Data System*, Planetary and Space Science, European Geophysical Society, ISSN 0032-0633, Volume 44, Number 1, January, 1996.

[2] Crichton, D. Hughes, J.S. ; Hardman, S. ; Law, E. ; Beebe, R. ; Morgan, T.; Grayzeck, E. A Scalable Planetary Science Information Architecture for Big Science Data. IEEE 10th International Conference on e-Science, October 2014.

[3] *Vision and Voyages: Vision and Voyages for Planetary Science in the Decade 2013-2022*, National Research Council, 2011.

LARGE SCALE DATA MANAGEMENT OF ASTRONOMICAL SURVEYS WITH ASTROSPARK

Mariem Brahem, Karine Zeitouni, and Laurent Yeh

DAVID lab., Univ. Versailles St Quentin,
Versailles France, Paris Saclay University

ABSTRACT

Large scale sky surveys has become a prominent topic in astronomy. The use of new telescopes with wide fields of view has enabled these surveys to deliver huge datasets. As a result, it's critical to provide a new framework that supports scalable and high performance query processing of these datasets. Apache Spark has been widely adopted as a successor to Apache Hadoop MapReduce to analyze Big Data in distributed frameworks. Despite its rich features, this framework can not be directly exploited towards processing astronomical data. In this work, we present AstroSpark, a distributed data server for astronomical data. AstroSpark extends Spark, a distributed in-memory computing framework, to analyze and query huge volume of astronomical data. It supports astronomical operations such as cone search, cross-match and k nearest neighbor (kNN). AstroSpark introduces effective methods for efficient astronomical query execution on Spark through data partitioning with HEALPix and customized optimizer. Experiments have shown that AstroSpark is effective in processing astronomical data, scalable and overperforms the state-of-the-art.

Index Terms— Astronomical Survey Data Management, Big Data, Query Processing, Spark Framework

1. INTRODUCTION

Large amounts of astronomical data are continuously collected, thanks to the improvement in the instrumentation side. For instance, the GAIA mission [3] and the future LSST survey are expected to produce petabytes of data. The analysis of such surveys is the basis of subsequent astronomical discoveries. For example, cross-matching enables astronomers to identify and correlate objects belonging to different observations in order to make new scientific achievements by studying the temporal evolution of the sources or combining physical properties. Meanwhile, such analysis is data-intensive since it involves access to billions of objects. Also, most astronomical queries are very expensive to process because of their compute-intensive nature especially for complex operations like cross-matching. In this respect, the growing scale of observed surveys coupled with the compute-intensive nature of astronomical operations has become a challenge.

DBMS technologies. Solutions [13] [14] [15] [16] based on relational DBMS have been proposed for querying astronomical datasets. However, these systems are based on centralized server architecture and even though some use multiple parallel disks, they could not provide a scalable solution.

Distributed frameworks. Recently, the shared-nothing type of parallel architecture, which uses commodity hardware, is becoming a de facto standard in big data handling. In this context, the distributed in-memory computing framework Apache Spark has emerged as a fast and general engine for large-scale data processing [12]. While Spark fits well the large scale nature of astronomical data, it does not provide native support of astronomical queries. But, users can rely on UDFs to process large astronomical data. For example, implementing a UDF to execute a cross-matching query leads to an expensive query evaluation through a cartesian product. In addition, astronomical queries require effective access methods to reduce the query search space and load only target partitions.

Spatial distributed frameworks. There exists a number of systems that support spatial data over distributed frameworks [7] [10] [11] [9]. However, these systems suffer from these limitations:

- The lack of a high level query language adapted to the astronomical context (like ADQL)
- The proposed operations are not adapted to the spherical coordinates system which leads to erroneous query results.
- Performance limitations in query processing due to the increase of objects along partitions borders.

Thus, there is a need to redesign and adapt existing distributed framework like Spark to support astronomical data.

In this work, we present AstroSpark [6], a distributed data server tailored for astronomical data. AstroSpark bridges the gap between the existing approaches listed above, we introduce a new framework for the management of large volume of astronomical data. AstroSpark is designed as an extension of Apache Spark that takes into account the peculiarities of the data and the queries in cosmological applications. Queries are expressed in Astronomical Data Query Language (ADQL)

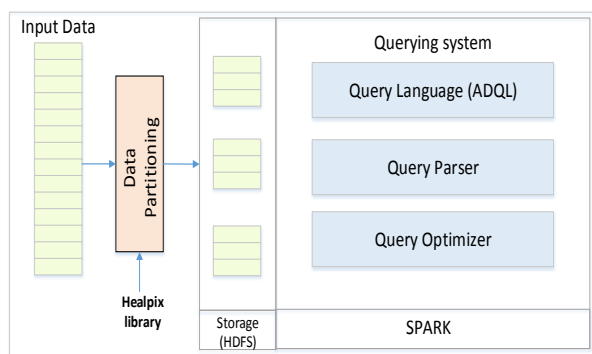


Fig. 1: AstroSpark Architecture.

[1], an SQL extension with astronomical functions. Various logical and physical optimization techniques for the ADQL execution are proposed, and integrated to Spark SQL thanks to the extensibility of its optimizer. This includes the control of the data partitioning mechanism and spatial indexing. AstroSpark implements a partitioner that achieves both spatial locality and load balancing. It also adopts a well-known sky pixelization technique, HEALPix (Hierarchical Equal Area isoLatitude Pixelization) [8] to organize and efficiently access the data. AstroSpark is under development. So far, the proposed algorithms, e.g., cone search, cross match, have shown drastic improvement of the performance by comparison with the state of the art [5].

2. ASTROSPARK PROPOSAL

AstroSpark is a distributed data server for Big Data in astronomy. It is based on Spark, a distributed in-memory computing framework, to analyze and query huge volume of astronomical data. AstroSpark in a nutshell adapts data partitioning to efficiently processing astronomical queries. To this end, we apply a spatial-aware data partitioning, and first use linearization with the HEALPix library to transform a two dimensional data points (represented by spherical coordinates) into a single dimension value represented by a pixel identifier (HEALPix ID). We have adopted the HEALPix approach for the following reasons:

- It is a mapping technique adapted to the spherical space, with equal areas per cell all over the sky, a unique identifier is associated to each cell of the sky. This id is a perfect index for astronomical sources and objects.
- Data linearization with HEALPix ensures preserving data locality, that is neighboring points in the two dimensional space are likely to be close in the corresponding one dimensional space. Data locality helps us to organize spatially close points in the same partition or in

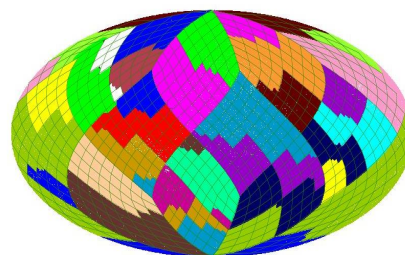


Fig. 2: Partitioning in AstroSpark

consecutive partitions, and thus optimizes query execution by reducing accesses to irrelevant partitions.

- The HEALPix library is maintained by the NASA. It is easily accessible and contains many functionalities that are useful in our context such as filtering neighbor pixels of those in a cone.

AstroSpark architecture is represented in figure 1. Queries are expressed using ADQL [1]. The query parser of AstroSpark is extended to translate an ADQL query with astronomical functions and predicates into an internal algebraic representation. Then, the query optimizer will enrich some pre-filtering operators based on our spatial partitioning which make global filtering prune out irrelevant partitions. AstroSpark extends the Spark SQL optimizer called Catalyst by integrating particular logical and physical optimization techniques. The interested reader may refer to [6] for more details.

2.1. Data Partitioning

Partitioning is a fundamental component for processing in parallel. It reduces computer resources when only a sub-part of relevant data are involved in a query, and then improve query performances. AstroSpark Partitioner should ensure two main requirements: (1) Data locality: points that are located close to each other should be in the same partition, a partition has to represent a portion of the sky (2) Load balancing: the partitions should be roughly of the same size to efficiently distribute tasks between nodes of the cluster. To achieve the first requirement, a spatial grouping of the data is necessary. Nevertheless, a basic spatial partitioning may lead to imbalanced partitions due to the typical skewness of astronomical data. Therefore, the partitioning should be also adaptive to the data distribution. AstroSpark partitions Spark dataFrames in a way the partitions are balanced while favoring data locality. We have visualised the created partitions using Aladin [4], a tool for viewing astronomical data and acquiring sky maps. (see figure 2).

The two dimensional spherical coordinates are first mapped into a single dimensional ID using the HEALPix library, which fulfil the data locality requirement. To achieve the load bal-

ancing, we leverage the Spark range partitioner based on this ID, which yields data partitions with roughly equal sizes. Then, the partitions are stored on HDFS. Each partition is saved in a separate subdirectory containing records with the same partition number. AstroSpark divides each partition into buckets. This technique optimizes query execution in a way that makes it efficient to retrieve the contents of a bucket and obviate scanning irrelevant partitions. AstroSpark retrieves also partition boundaries and store them as metadata. Note that in our case, all we need to store are the three values (n, l, u) where n is the partition number, l is the first HEALPix cell of the partition number n and u is the last HEALPix cell of the partition number n . It should be noted that we store the partitioned files (with the metadata) on HDFS and use them for future queries, which amortizes their construction cost.

2.2. Query Parsing

The query parser allows the extension of Spark to deal with astronomical queries and ADQL. We have proposed and evaluated three alternatives of query parsing:

- Case 1: DataFrame API. If the query is written using the DataFrame API, we provide a query interface by extending this API with astronomical operations.
- Case 2: Query rewriting. If the query is written using ADQL, we replace all parts of the query matching an astronomical pattern with replacement SQL text. The query parser verifies first that the query is syntactically correct. Then, it extracts tables names, columns name and some keywords such as CONTAINS, JOIN, POINT, CIRCLE from the input query. This extraction helps AstroSpark to fetch the query type. After that, we apply some transformations to the original query to optimize it.
- Case 3: Strategies. Using this alternative, we start by checking the ADQL query to make sure that the syntax is valid. If the query syntax is correct, the query parser converts the ADQL query in a logical plan which is then transformed into an optimized physical plan using our customized strategies. Query parsing performed in AstroSpark is completely transparent to the user. However, for the time being, our query parser does not support all the ADQL syntax.

AstroSpark provides two query interface (ADQL and DataFrame API) to execute specific astronomical operations (cone search, cross-match and kNN). To support generic ADQL grammar, users need to add some UDFs which have already been implemented in other works [2].

2.3. Query Optimization

The query optimizer is an important component since it is the responsible for generating a query execution plan that

computes the query result efficiently. AstroSpark extends the Spark Catalyst optimizer by adding custom strategies for astronomical queries. We integrate new strategies for converting Spark naive logical plan to an optimized physical plan. The output physical plan is the actual plan which AstroSpark executes for the final data processing. The query optimizer performs the implemented strategies on the input query to transform the query tree into equivalent, but with optimized form.

3. RELATED WORK

Recent works have addressed the support of spatial data and queries using a distributed data server. Their architecture have followed the development of the Hadoop ecosystem. We devise three representative proposals.

SpatialHadoop [7] is an extension to Hadoop that supports spatial data types and operations. It improves each Hadoop layer by adding spatial primitives. SpatialHadoop adopts a layered design composed of four layers: language, storage, MapReduce, and operations layers. For the language layer, it adds an expressive high level SQL-like language for spatial data types and spatial operations. In the storage layer, SpatialHadoop adapts traditional spatial index structures, Grid, R-tree and R⁺-tree, to form a two-level index structure of global and local indexing. SpatialHadoop enriches the MapReduce layer by adding two new components, SpatialFileSplitter and SpatialRecordReader. In the operations layer, SpatialHadoop focuses on three basic operations range query, spatial join, and k nearest neighbor (kNN).

MD-HBase [10] is a scalable multi-dimensional data store for Location Based Services (LBSs), built as an extension of HBase. MD-HBase supports a multi-dimensional index structure over a range partitioned Key-value store, builds standard index structures like k-d trees and Quad-trees to support range and kNN queries.

GeoSpark [11] extends the core of Apache Spark to support spatial data types, indexes, and operations. In other words, the system extends the resilient distributed datasets (RDDs) concept to support spatial data. GeoSpark provides native support for spatial data indexing (R-Tree and Quad-Tree) and query processing algorithms (range queries, kNN queries, and spatial joins over SRDDs) to analyze spatial data.

More recently, SIMBA [9] is an extension of SPARK-SQL (not only at the core level of Spark) to support spatial queries and analytics over big spatial data. SIMBA builds spatial indexes over RDDs. It offers a programming interface to execute spatial queries (range queries, circle range queries, kNN, Distance join, kNN join), and uses cost based optimization.

The aforementioned systems are designed for the geo-spatial context that differs from the astronomical context in its data types and operations. These systems (except SIMBA) deal only with spatial data types and do not support the combina-

tion of additional attributes and relational operators. They do not provide a high level query language adapted to the astronomical context like ADQL, and do not offer astronomical functions which are tailored for the spherical coordinates system. The astronomical context uses also specific operations such as cone search queries, cross-match queries, and kNN queries that are not supported by these systems.

In the astronomical field, the VizieR service [13] developed by the Centre de Données de Strasbourg (CDS), is available to provide a tool for accessing astronomical data listed in published catalogs and executing cross-matching queries. In addition, some recent works [14] [15] [16] proposed customized solutions to execute main astronomical queries. In [14], authors introduce a cross-matching function including a partitioning approach with the HEALPix library. Q3C [15], Quad Tree Cube provides a sky partitioning schema for PostgreSQL and offers an SQL interface for main astronomical queries. OPEN SKYQUERY [16] allows querying and cross-matching distributed astronomical datasets. The authors propose zoning and partitioning algorithm to execute queries and facilitate parallelization using RDBMS technologies. However, none of these efforts provides a scalable astronomical server with a unified programming interface like ADQL, as we target. Their performances are still limited because they are mainly based on centralized server architecture style.

4. CONCLUSION

In this paper, we described AstroSpark, a distributed in-memory engine for astronomical queries. Our framework extends Spark in order to allow astronomers to process astronomical data using the expressivity of ADQL. Experiments on real datasets from the on-going spatial mission *GAIA* demonstrate that AstroSpark is much faster than others systems.

5. ACKNOWLEDGMENTS

This work is partly funded by the Centre National d'Etudes Spatiales (CNES). It has made use of data from the European Space Agency (ESA) mission *Gaia*, processed by the *Gaia* Data Processing and Analysis Consortium DPAC. Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

6. REFERENCES

- [1] ADQL. <http://www.ivoa.net/documents/latest/ADQL.html>.
- [2] ADQL CDS. <http://cdsportal.u-strasbg.fr/adqltuto/>
- [3] *GAIA*. <http://sci.esa.int/gaia/>
- [4] ALADIN. <http://aladin.u-strasbg.fr/>
- [5] Brahem, Mariem, Karine Zeitouni, and Laurent Yeh. "HX-MATCH: In-Memory Cross-Matching Algorithm for Astronomical Big Data." International Symposium on Spatial and Temporal Databases. Springer, Cham, 2017.
- [6] Brahem, Mariem, et al. "AstroSpark: towards a distributed data server for big data in astronomy." Proceedings of the 3rd ACM SIGSPATIAL PhD Symposium. ACM, 2016.
- [7] Eldawy, Ahmed, and Mohamed F. Mokbel. "Spatial-Hadoop: A MapReduce framework for spatial data." Data Engineering (ICDE), 2015 IEEE 31st International Conference on. IEEE, 2015.
- [8] Gorski, Krzysztof M., et al. "HEALPix: a framework for high-resolution discretization and fast analysis of data distributed on the sphere." The Astrophysical Journal 622.2 (2005): 759.
- [9] Xie, Dong, et al. "Simba: Efficient in-memory spatial analytics." Proceedings of the 2016 International Conference on Management of Data. ACM, 2016.
- [10] Nishimura, Shoji, et al. "MD-HBase: design and implementation of an elastic data infrastructure for cloud-scale location services." Distributed and Parallel Databases 31.2 (2013): 289-319.
- [11] Yu, Jia, Jinxuan Wu, and Mohamed Sarwat. "Geospark: A cluster computing framework for processing large-scale spatial data." Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2015.
- [12] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." HotCloud 10.10-10 (2010): 95.
- [13] Ochsenbein, Francois, Patricia Bauer, and James Marcout. "The VizieR database of astronomical catalogues." Astronomy and Astrophysics Supplement Series 143.1 (2000): 23-32.
- [14] Zhao, Qing, et al. "A paralleled large-scale astronomical cross-matching function." Algorithms and Architectures for Parallel Processing (2009): 604-614.
- [15] Koposov, S., and O. Bartunov. "Q3C, Quad Tree Cube—the new sky-indexing concept for huge astronomical catalogues and its realization for main astronomical queries (cone search and Xmatch) in open source database PostgreSQL." Astronomical Data Analysis Software and Systems XV. Vol. 351. 2006.
- [16] Nieto-Santisteban, Mara A., Aniruddha R. Thakar, and Alexander S. Szalay. "Cross-matching very large datasets." National Science and Technology Council (NSTC) NASA Conference. 2007.

CONSIDERING SCALE OUT ALTERNATIVES FOR BIG DATA VOLUME DATABASES WITH POSTGRESQL

Pilar de Teodoro, Sara Nieto, Jesus Salgado, Christophe Arviset

European Space Astronomy Centre, Madrid, Spain

ABSTRACT

When a new mission is planned, its archive must also be planned. Several years of the mission development will also lead to a developed archive system that will be finally the data remaining from the mission.

In this paper a study of the evolution of the databases holding the data from different archives will be presented from the ESAC Science Data Centre (ESDC) perspective and especially the scale out tests performed in PostgreSQL databases for holding the new generation of archives, which will hold terabytes to petabytes of data.

Index Terms— databases, archives, scale-out, PostgreSQL

1. INTRODUCTION

Traditionally the data belonging to a mission archive is divided in data located in a file system and catalogues and metadata stored in a database software provider system. The data could come at the end of a mission or just being part of the mission, collecting significant data that will serve scientist from the mission consortium first to analyze the mission data and later to the public.

The structure of this archive system will depend on the requirements from the mission normally evolved from the continuous work between the archive scientists and the development teams from the archive group. But the challenge of the new archives is the big amount of data generated from the missions that must be offered in a quick and useful way to the scientists.

At the ESAC Science Data Centre the engineers and scientists work for creating the Astronomy, Planetary and Helio mission archives from ESA missions and the continuous challenge is how to work with the demanding big data volume issues.

2. DATA STORAGE AT THE ARCHIVES

The systems storing the data had evolved tremendously to be stored first all data on a single machine and then, moved to Network Attached Storage systems and/or machines with local terabytes of Solid State Disks, some with flash memory included to enhance the performance in a significant way.

The following graph (Figure 1), presents the size of

each ESDC mission archive PostgreSQL database. The databases hold primarily metadata describing the scientific contents of the data stored in files repositories.

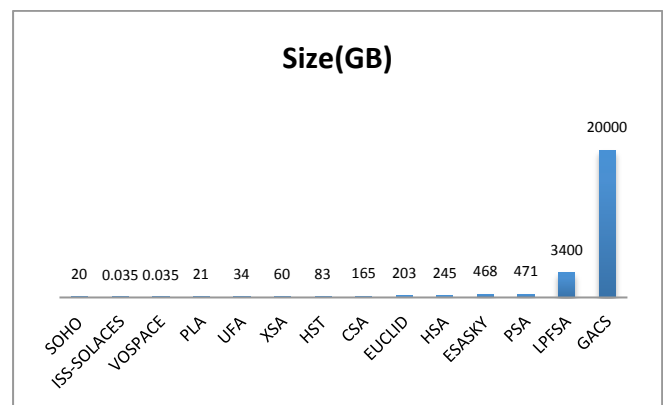


Figure 1: Database size per mission archive

At the time of writing this paper the archive from Gaia (GACS) [1] is the largest of our databases with 20 terabytes of data, it contains catalogues, user spaces and auxiliary data, followed by Lisa Pathfinder archive (LPFSA) with 3,4 terabytes of which more than 3 terabytes are telemetry data from the different mission phases that can be queried directly from the archive application.

Gaia Data Release 2 (DR2), will contain a bigger catalogue than DR1[2] which includes time series data and will be served from a PostgreSQL[3] scale-out cluster, which paves the way of using scale-out solutions for the archives. Due to the requirements of this kind of archives of once ingested- many times read, it serves as a very suitable solution.

The Euclid mission [4] [5] will be much larger than Gaia with an expected total size of about 175 petabytes of which a 3 to 5% is expected to be a central metadata repository database for running the pipelines and a scientific archive system database which will support the catalogues, user spaces and scientific metadata releases. The evolution in time of the ESDC archives including file repositories are shown in figure 2, Euclid as launched by Q3 of 2021 will be by 2025 the largest of the ESDC catalogues and will even grow more towards the end of the mission. This introduces the need of finding other solutions than the ones deployed so far such as monolithic RDBMS databases implementations.

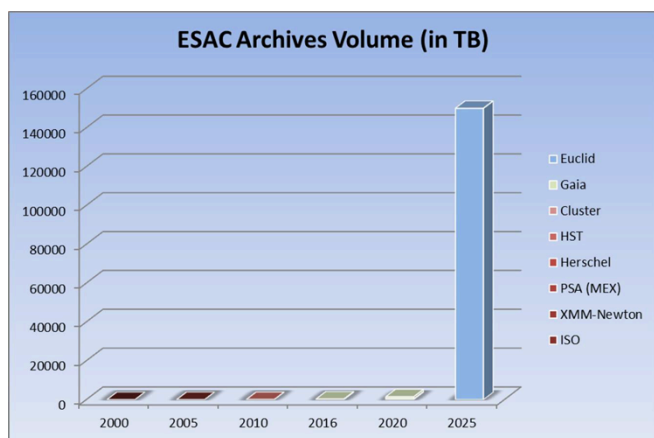


Figure 2: Evolution of archives repositories

3. ACCESSING THE DATA

As the archives visible door is a web portal (Figure 3) allowing the users to query the different archives, it is natural to store the metadata and catalogue data in a database for fast access to the data. The main website for the archives is: <http://archives.esac.esa.int>.

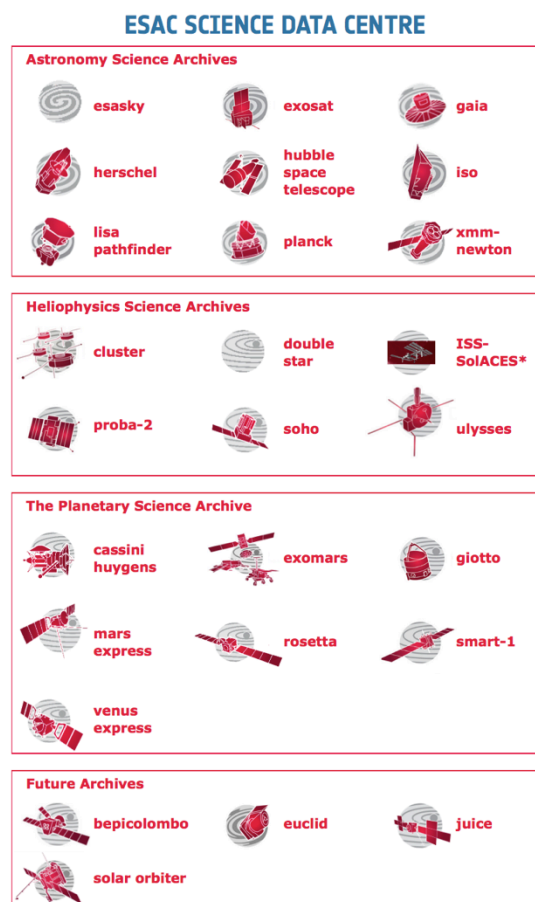


Figure 3: ESDC archives Web Portal

4. DATABASE SYSTEMS

Currently the main database system used is an open source system, PostgreSQL [3], very popular in the scientific community with many contributors and which holds excellent extensions for spherical queries.

Postgres or PostgreSQL has been proved as a key component of the ESA Archives allowing not only spherical queries but also the use of healpix, q3c and postgis extensions for geographical indexing and analyzing time series data with timescaledb extension. High availability and fault tolerance is another advantage of this system. Postgres is a RDBMS but as the data volume grows it can evolve in a scale out manner.

5. CONTEXT

Gaia mission DR1 [1][2] is currently the ESDC largest archive system with 20 terabytes of data, it contains the Gaia Source catalogue which is 1,5 terabytes in size including 1,2 terabytes of indexes defined for faster access. Those indexes so far can fit in memory in a single machine but for next DR3 with hundreds of terabytes of data expected and hundred of terabytes of indexes to be created, the scale up strategy will not be enough for fast access and other possibilities need to be covered.

The same situations will need also to be covered for the Euclid mission [4][5], Solar Orbiter mission and future missions. In this context, and having the experience of working with PostgreSQL, different scale out options have been studied. These postgres flavours rely in a distributed system that can remind in a certain way to a distributed Hadoop architecture system.

6. SCALE OUT POSTGRESQL OPTIONS

Currently in the market there exist 3 options for distributed PostgreSQL architecture for scaling out:

- 2nd Quadrant Postgres-XL [7]
- CitusData [8]
- OpenSource GreenPlum [9]

For resources reasons, the scale out options tested were Postgres-XL and Citusdata. GreenPlum has recently opened source and will be considered in a future. We will present the results of the two studies that will provide the experience of choosing the best options depending on the archive needs.

7. TESTS

7.1. Postgres-XL

Postgres-XL is a horizontally scalable open source SQL database cluster, flexible enough to handle varying database workloads. The tests done to prove scalability rely on the following architecture (Figure 4):

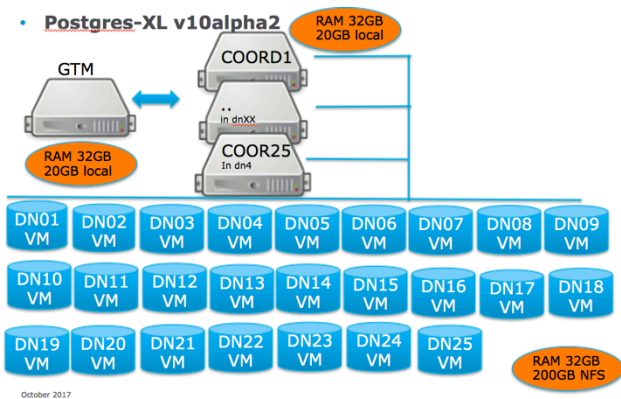


Figure 4: Postgres-XL test setup

Connections to the distributed database are done based on a connection to one of the coordinators. Having several coordinators allows to parallelized ingestion and queries when using a connection manager such as HAProxy. The Global transaction manager keeps the logic for connecting the information between the datanodes. Several tests have been done with the KiDS dataset [6] using 1 single instance, 3, 6,10,16 and 25 nodes to show scalability for simple queries and aggregates functions.

7.1.1. Aggregate functions tests

The scalability has been proved for aggregate functions as can be seen below (Figure 5) when plotting for a simple query:

```
select count(*) from kids_mb_catalogue;
```

Time in milliseconds versus number of rows retrieved shows a difference of 20 times faster for a distributed query in 25 nodes compare with the results in a single instance.

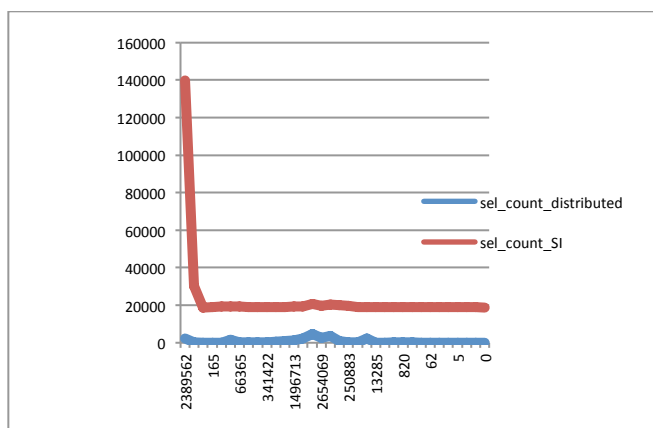


Figure 5: aggregate functions

7.1.2. Simple queries Tests

The solution scales out for simple queries such as: `select * from kids_mb_catalogue where mag_auto_i=XX`. The results are written in a single file with the output of each query. The tests performed are summarized in figure 6.

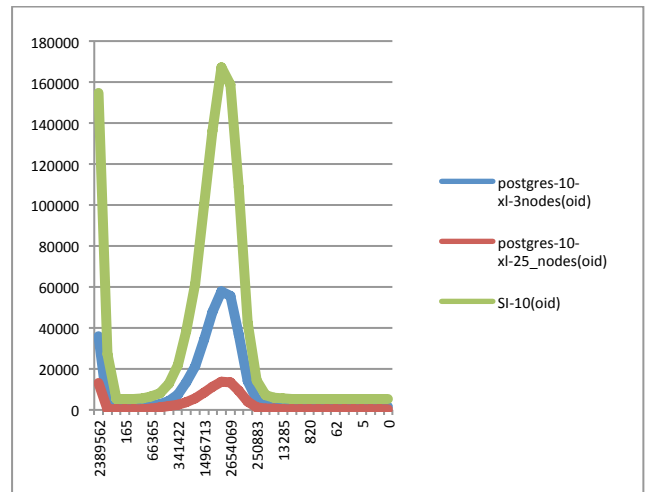


Figure 6: Comparison of scalability, time in ms vs Rows number

It is clear that using 25 datanodes is much faster than a single instance, as the query is splitted in those 25 nodes. The coordinator only needs to append the output.

7.2. CitusData

Citus [7] is a distributed database that extends PostgreSQL, allowing you to continue using all the powerful Postgres features while still scaling.

The main difference with postgres-xl is that it is an extension of postgres and not a fork so it can support other extensions such as foreign data wrappers, which are not supported on postgres-xl so far. There is no GTM node as the logic resides on the master node. By default the master node is only one and it needs to be replicated in case the setup wants to benefit for parallel ingestion using several master nodes. The data shard factor can be increased (default is 1) to allow having replicas of the data in the worker nodes or data nodes, in the same way a High Distributed File System (HDFS) works. The configuratsetup used for the tests are described in figure 7.

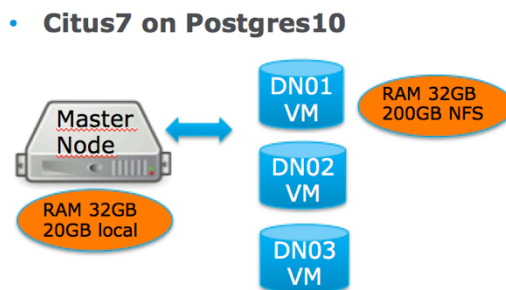


Figure 7: Citius setup

Several tests were run with different version of Citus, 6 and 7, using a postgres 9.6 and postgres-10 beta4 versions.

In the following plot (figure 8), Citus 7 was tested using a postgres 9.6 and a postgres 10 release with different sharding keys. One test using OID as unique identifier by each row in the table as primary key with a distribution made by hash and other tests using multitenant_id partition based on a truncated value of mag_auto_i table which holds values from 10 to 43.

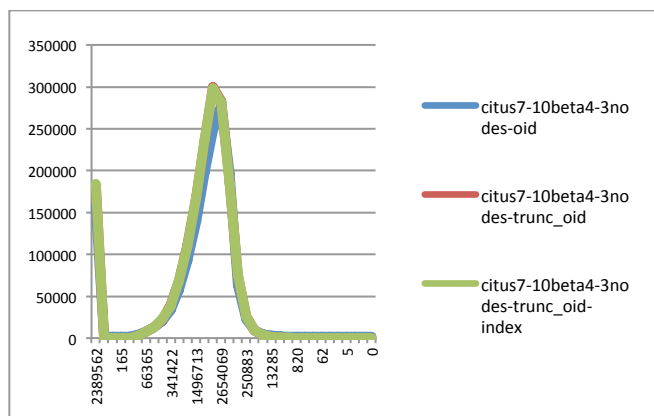


Figure 8: Citus 7 tests: Time in ms vs number of rows retrieved by filter

In CitusData we have not seen great improvements using the multitenant sharding vs hash sharding, it is slightly better the hash distribution but even using an index on the filter is not used in the execution plan.

7.3. Comparison between Postgres-XL and Citus:

As a summary we have plotted (figure 9) the performance of the simple query using a Single Instance on postgres 10 beta4, Postgres-XL solution running in 3 and 25 datanodes and Citus7 with a postgres 10 with oid hash distribution. Postgres-XL in 25 nodes is x10 faster than Citus in this kind of tests.

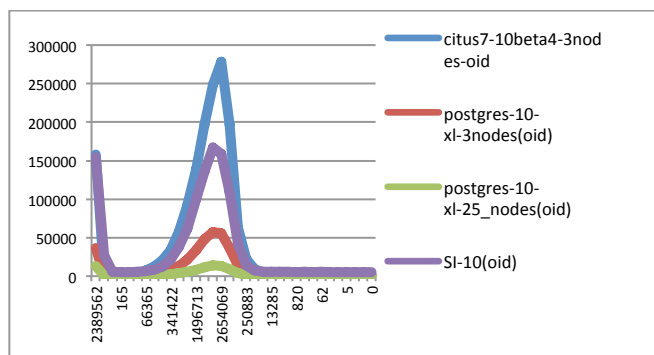


Figure 9: Citus 7 vs postgres-xl in 3, 25 nodes and single instance

8. CONCLUSIONS

The evolution of the missions requires new investigation on data storage and management. A single option does not fit all cases and choosing the right one for each archive will be necessary. Scaling out is a must for the future archives and we want to continue relying on open source software. The Gaia archive will be the first on using a scale-out solution from Postgres-XL, the Euclid archive will follow. More tests will be needed to ensure high availability, robustness, and good performance for ingestion and querying. Time to administer the resources and support from the community and the provider is also important. We have to get a balanced solution between hardware and software to fulfill the requirements for each mission.

9. REFERENCES

- [1] C. Arviset. “Big data, big challenges and new paradigm for the Gaia archive”. Proceedings for the BIDS conference. Santa Cruz de Tenerife. March 2016. p.9.[LBNA27775ENN.pdf](#)
- [2] T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, C. Babusiaux, C. A. L. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans and et al. (2016b) “The Gaia mission”. External Links: 1609.04153
- [3] PostgreSQL database: <http://www.postgresql.org>
- [4] P. Teodoro, S.Nieto. “The Euclid Archive System, a data centric approach to Big Data”. Astrominformatics: proceedings of the 325th symposium of the International Astronomical Union held in Sorrento, Italy, October 19-25 2016 / edited by Massimo Brescia [and four others]. ISBN: 110716995X
- [5] M. Poncet . Euclid: “Big Data from Dark Space” Science Ground Segment Challenges for next decade. Proc. of the 2014 conference on Big Data from Space (BiDS’14) p.167
- [6] Jelte T. A. de Jong, Gijs A. Verdoes Kleijn, Konrad H. Kuijken, Edwin A. Valentijn, KiDS and Astro-WISE consortiums. “The Kilo-Degree Survey. “ June 6, 2012 External link: <https://arxiv.org/pdf/1206.1254.pdf>
- [7] Postgres-XL:<http://www.postgres-xl.org>
- [8] CitusData:<https://www.citusdata.com>
- [9] Open Source GreenPlum: <http://greenplum.org>

ARCHIVE MANAGEMENT OF NASA EARTH OBSERVATION DATA TO SUPPORT CLOUD ANALYSIS

Christopher Lynnes, Kathleen Baynes, Mark McInerney

National Aeronautics and Space Administration

ABSTRACT

NASA collects, processes and distributes petabytes of Earth Observation (EO) data from satellites, aircraft, in situ instruments and model output, with an order of magnitude increase expected by 2024. Cloud-based web object storage (WOS) of these data can simplify the execution of such an increase. More importantly, it can also facilitate user analysis of those volumes by making the data available to the massively parallel computing power in the cloud. However, storing EO data in cloud WOS has a ripple effect throughout the NASA archive system with unexpected challenges and opportunities. One challenge is modifying data servicing software (such as Web Coverage Service servers) to access and subset data that are no longer on a directly accessible file system, but rather in cloud WOS. Opportunities include refactoring of the archive software to a cloud-native architecture; virtualizing data products by computing on demand; and reorganizing data to be more analysis-friendly.

Index Terms— Cloud computing, archive, architecture, data analytics

1. INTRODUCTION

NASA collects and processes increasingly large volumes of Earth Observation (EO) data from satellites, aircraft, in situ instruments and model output. NASA's Earth Observing System Data and Information System (EOSDIS) is responsible for archiving the data and distributing them to a variety of end user communities, including science researchers and applied science users [1]. EOSDIS EO data archives comprise 12 Distributed Active Archive Centers at a variety of locations in the United States who save the data mostly on on-premise disk arrays, with some tape storage. These archives are knitted together by a Common Metadata Repository of metadata at the data collection and file level, which allows a search client to search across all 12 DAACs using a single database.

Since the turn of the century, the data volume archived in has increased 400-fold, to approximately 25 PB in 2017. An additional order of magnitude increase is expected by the year 2024 (Fig 1). Just as important, EOSDIS distributes an annual data volume that is of the same order of magnitude as its cumulative archive volume.

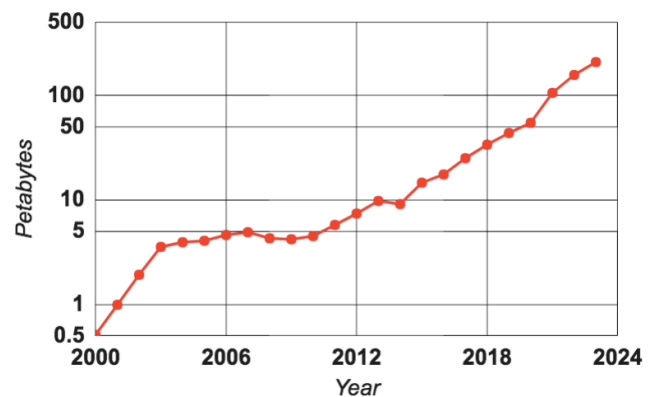


FIGURE 1. HISTORICAL AND PROJECTED CUMULATIVE ARCHIVE VOLUME IN EOSDIS. (YEARS RUN FROM OCTOBER TO SEPTEMBER.)

Cloud-based storage simplifies the ramp-up to handle such large volume increases. It obviates the need to specify and procure large amounts of hardware, plus many of the ancillary activities required, such as allocating (or build out) precious raised floor space, tracking property items, upgrading cooling systems, and upgrading internal networks. Also, the diversity of the community served by major cloud vendors has led to a variety of storage options with different latency, throughput and access options balanced against the respective costs.

However, while modest cost savings may be achievable by using cloud storage over on-premise storage, the real potential arises in the proximity of enormous computing power “next to” the cloud storage. In theory, science researchers using the data could now apply data-parallel processing to analyze data volumes that would simply be too big to download and analyze with their own hardware. Another potential advantage is that having so many data collections in one virtual “place” lends itself to more multi-data-collection studies; these are a particular feature of Earth Observation studies which often meld data from multiple sources based on satellites, aircraft, in situ and model outputs. In reality, the data are often physically separate in cloud storage, but the high-bandwidth interconnects within clouds mitigate this distance.

2. ARCHIVING EO DATA IN CLOUD STORAGE

For the above reasons, NASA is exploring a variety of prototypes using public cloud to archive and distribute EO data. Several of the prototypes use Web Object Storage (WOS) for data archiving in the form of Amazon Web Services Simple Scalable Storage (AWS S3). The prototypes identify the business and operational implications of archiving data in the cloud, as well as demonstrating some of the potential benefits from cloud-based archives.

The core prototype in this suite, named Cumulus, is developing a science archive hosted in the public cloud. Rather than lift and shift an existing archive software system from within EOSDIS, a conscious decision was made to employ cloud-native architecture and services in the prototype. This enabled an architecture centered on AWS Lambda functions triggered by the arrival of data notices, and orchestrated through Step Function workflows. As a result, the workflow aspect is handled largely by cloud-provided services, with the result that most of the custom code is focused on the “business logic”, in this case the ingest and processing of different EO science products. The ideal would be to have custom code only for the specific business logic, with cloud services supplying the software infrastructure.

One impact of storing the data into S3 is the egress cost of distributing data out of the cloud, which is exacerbated by the short-term uncertainty of user-requested egress. A traffic rate shaper can mitigate this, taking care to not impact user access unduly. On the other hand, transferring data from WOS to a compute node in the same cloud region does not incur egress cost, incentivizing users to make the paradigm shift toward analyzing data in place (or nearby), rather than downloading to a local machine.

This also incentivizes the archive to offer a number of data reduction services to aid the user in preprocessing the data and decreasing the volume that might be transferred out of the region. Currently, EOSDIS offers several subsetting services, particularly based on the Open-Source Project for a Network Data Access Protocol (OPeNDAP) [2]. Other services to support custom subsetting, regridding, reprojection, quality screening and mosaicking are offered for certain data products. Most of these services are designed to run on a host with attached storage in the form of a POSIX filesystem. Instead, Web Object Storage offers data through the Hypertext Transfer Protocol. Thus, one of the prototypes employs a form of OPeNDAP server that can serve data in Web Object Storage.

Another novel aspect of archiving data in the cloud is that costs accrue so long as the data rest in storage, which can add up to significant expenditures over time. One approach to address this is to virtualize some of the high-volume data and produce them on demand. This strategy was previously employed in serving MODIS Calibrated Radiance data during the transition of EOSDIS from mostly tape to mostly disk in the mid 2000’s. However, once disk prices had dropped enough to afford to put the MODIS Calibrated Radiance on disk, this strategy was largely phased out, due to the relatively large latency of on-demand production to simply serving from disk. However, while large on demand requests may be painfully slow for big requests on current computer systems, the ability to access hundreds or thousands of compute nodes at once in the cloud could conceivably shrink the response time to be almost indistinguishable from serving the data from storage. At that point, it becomes a tradeoff between the cost of compute cycles needed to make virtual products vs. the storage cost.

The broad ecosystem of cloud services to fulfill common functions provides another opportunity. In the course of prototyping, we can refactor the archiving software system to use off the shelf services, such as queues, databases, and workflow support services to dramatically reduce the code base.

3. SUPPORTING CLOUD ANALYTICS

Ultimately, the “killer app” for archiving in the cloud is to support analytics using the massively parallel capabilities offered by cloud computing. This has a particularly wide variety of solutions being explored in the community. Most of them involve sharding data across a large number of nodes to enable parallel computing. The sharding solutions include highly distributed databases (e.g., Cassandra [3]), highly distributed filesystems (e.g., Hadoop File System [4]) or in some cases simply dividing data up amongst many WOS buckets. These can be paired with an equally varied set of computational technologies (e.g., Spark [5]). Cloud prototypes are underway to develop end-to-end demonstrations of such systems, with three main aims: (1) to determine feasibility and operability; (2) to demonstrate to the science community what can be accomplished with cloud computing near the data; and (3) to determine the possible impacts on the archive architecture.

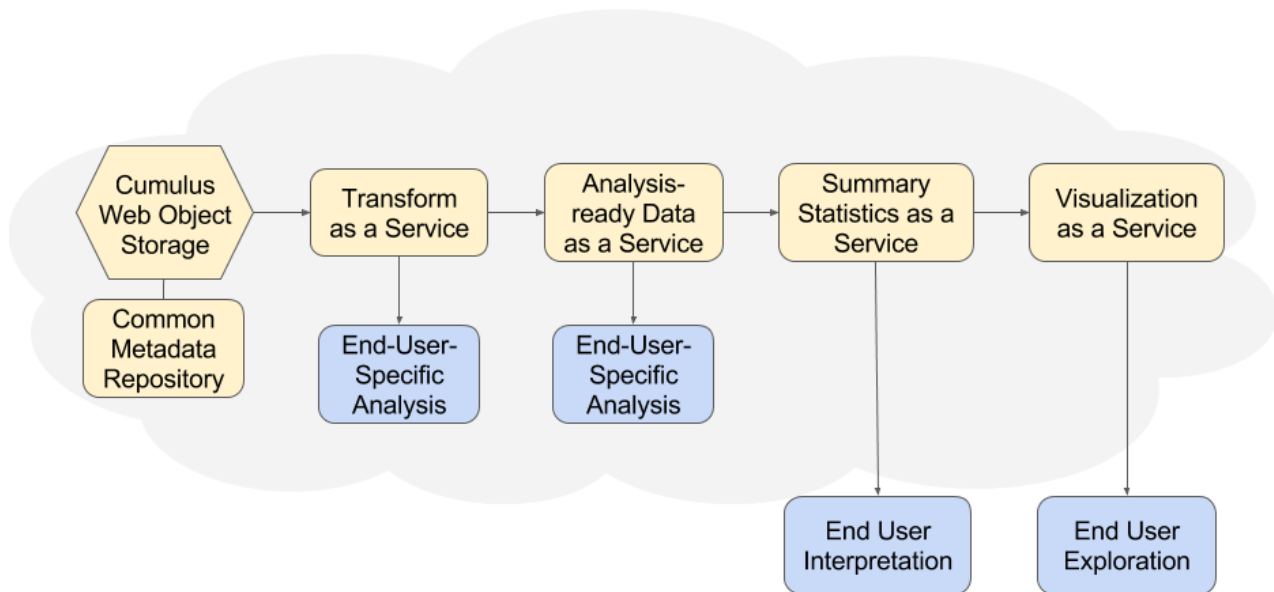


FIGURE 2. ABSTRACTED PIPELINE FOR DATA ANALYTICS IN THE CLOUD.

One common factor among most of these analysis technologies is that they usually require reorganizing and reformatting the data in order to store them in a highly distributed database or filesystem. These forms of analytics-optimized data storage typically have different performance characteristics when paired with appropriate corresponding analytics frameworks. Unfortunately, it is not yet clear if there is a universal optimum combination of analytics optimized storage and analytics frameworks with respect to cost and speed. The optimum may depend on the data characteristics, the analysis algorithm, and the user's specific use case, say, data exploration vs. in-depth analysis. Therefore, we are developing an architectural concept that abstracts the main steps in the analysis process, presenting them to the world as services: this will provide a common framework that can accommodate different components for different combinations of data and use cases (Fig. 2).

A typical end-to-end analysis begins with extraction of the necessary data variables for the spatial and temporal Region of Interest from the Cumulus Web Object Storage. A Common Metadata Repository stores the essential metadata that allow us to generalize this process to work with many types of data in EOSDIS. This is followed by optional data transformation steps, which may include quality filtering, regridding, and/or aggregation over time, space or variables. This corresponds to the "Transform" of the common Extract-Transform-Load process in analysis pipelines. The data are then stored in an analytics-optimized storage framework such as Parquet, HDFS, or Cassandra. The next step provides a set of summary statistics that are commonly used in the EOSDIS user community, usually involving an averaging over latitude, longitude, or time. Based on the remaining dimensions in the data, this is

followed by visualization. This overall flow is exemplified by the Geospatial Interactive Online Visualization AND aNalysis Infrastructure (Giovanni) [6], a current EOSDIS on-premise tool currently being ported to the cloud. Giovanni serves over 1700 data variables to a user base measured in the tens of thousands.

The abstracted architectural concept for archive-proximal analysis in Fig. 2 follows the cloud computing pattern of exposing each key element as a service. This produces several salutary effects. First and foremost, it makes for an open system, one that allows a variety of analysis system developers to plug into the system at any step in the process. Similarly, it can serve an even wider diversity of users than the monolithic on-premise analysis solutions. Interdisciplinary users, educational users, and applications users can work with the visualizations that provide data exploration capabilities with little user effort. On the other hand, research scientists who create and use built-to-purpose analysis can gain value from the data preprocessing and reorganization available via the Transform-as-a-Service and Analysis-Ready Data as a Service. Analysis Ready Data have been promoted by the Committee on Earth Observing Satellites as "*are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets*"[7]. Note also that the data volume generally is smaller on the right side of the pipeline, with summary statistics and their visualizations usually representing a small fraction of the original volume from which they were generated. Thus, it is still important for egress charge reasons for end users to be able to easily apply their own

analyses on transformed and analysis-ready data within the cloud. On the other hand, there is little penalty to distributing summary statistics and visualizations to users outside the cloud.

One challenge of this architecture is that archives are hesitant to abandon the data format as received from the provider, implying that they will likely manage two or more copies in different forms. However, it may be cost prohibitive to keep the reorganized version on fast, expensive storage needed for high performance indefinitely. This is particularly the case when a clear winner in price per performance for different data storage technologies is still up in the air. This implies that strategies and mechanisms will be needed for deciding which data to make available in the analytics optimized form, and for how long. These strategies need to be flexible enough to adapt to the ever-changing cost and capabilities on offer by commercial cloud providers.

4. CONCLUSIONS

The heightened interest in Big Data in the larger business community has spawned an increase in off-the-shelf services that are useful for managing and processing data. Managing Big Data in Earth Observation archives can benefit from adopting many of the resultant capabilities. There are many challenges in pivoting from storing data on on-premise hardware to storing them in the cloud. However, there are at least as many opportunities to leverage the co-location of massive processing power near enormous storage resources in order to perform science analysis on larger datasets than ever before, as well as faster than ever before. Recognizing the significant (but sometimes subtle)

differences in cloud archive management continues to drive prototype development in NASA Earth Science systems to explore the opportunities and mitigate the risks inherent in such a major evolutionary change in archive architecture.

5. REFERENCES

- [1] H.K. Ramapriyan, R. Pfister, and B. Weinstein, "An Overview of the EOS Data Distribution Systems", *Land Remote Sensing and Global Environmental Change*, Springer, New York, pp. 183-202, 2010.
- [2] P. Cornillon, J. Gallagher, and T. Sgouros. "OPeNDAP: accessing data in a distributed, heterogeneous environment," *Data Science Journal* **2**, pp. 164–174, 2003.
- [3] A. Lakshman and P. Malik. 2010. "Cassandra: a decentralized structured storage system," *SIGOPS Oper. Syst. Rev.* **44**, 2, pp. 35-40, 2010.
- [4] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies*, Incline Village, NV, pp. 1-10, 2010.
- [5] J. G. Shanahan and L. Dai. "Large Scale Distributed Data Science using Apache Spark." *Proc. 21st ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, pp. 2323-2324, 2015.
- [6] Liu, Z., and J. Acker. "Giovanni: The bridge between data and science, *Eos*, 98, doi:10.1029/2017EO079299. 2017.
- [7] CEOS Analysis Ready Data for Land (CARD4L) Description Document, accessed at: http://ceos.org/document_management/Meetings/Plenary/30/Documents/5.5_CEOS-CARD4L-Description_v.22.docx

ONLINE EARTH OBSERVATION DATA MANAGEMENT

Nicolas Weiland⁽¹⁾, Stephan Kiemle⁽¹⁾, Markus Kunze⁽¹⁾, Torben Keßler⁽²⁾

⁽¹⁾ German Aerospace Center (DLR), Earth Observation Center (EOC), Oberpfaffenhofen, Germany

⁽²⁾ Werum Software & Systems, Lüneburg, Germany

ABSTRACT

The ever growing amount of exploitable remote sensing data, novel online IT infrastructures and the increasing need of applications for rapid processing of large coverages and long time series in high resolution have pushed Earth Observation (EO) data centers to set up exploitation platforms with large online storage and processing capacities. This way, “users are brought to where the data are” instead of transferring large amounts of data to the users.

However, the online storage of petabytes of EO data demands a data management which is somewhat different but not less challenging than the data management tasks of archiving data centers. In this paper, we discuss online EO data management requirements and present a functional architectural concept for systems managing EO data in exploitation platforms. The German Aerospace Center DLR applies this concept in the context of various projects and EO missions.

Index Terms— Data Management, Exploitation Platform, Online Storage

1. INTRODUCTION

Satellite-based Earth Observation systems have observed a major paradigm shift with the Copernicus era of satellites with high data rates, making it nearly impossible to systematically provide all payload data and processed products to any users’ premises. Network capacities are an expensive resource and users often do not want to set up and maintain large local storage infrastructures.

The paradigm shift in EO data exploitation is on the other hand pushed by the trend of EO applications needing to process larger amounts of data in shorter time. Various Earth sciences focus on change detection, working on long EO data time series analyzing phenomena at a global scale which require high resolution global data coverages. Iteratively enhanced algorithms allow rapid progress on information quality and on discovering new properties and interrelations. The same applies to institutional and commercial applications taking advantage of instantaneous access to large amounts of input data for adding value and extracting thematic information.

These trends are supported by evolving technologies of hosted IT infrastructures. Private, public, and combined

private-public cloud solutions promise flexible setup and dynamic allocation of storage and processing resources at the infrastructure level.

One answer to this shifting EO data access paradigm is the emergence of EO data exploitation platforms. Initially these were designed to purely provide dynamic processing resources, the users being responsible to provide mature processing software, to reference input and auxiliary data, and to store the processing results. In a next step, the platforms were extended to serve specific missions or specific thematic application domains, adding input and auxiliary data holdings, providing maintained data processors and tools, and supporting documentation and expert knowledge sharing.

2. EXPLOITATION PLATFORMS

Since Copernicus data have to be made available to a broad user community large multi-purpose access and exploitation platforms have emerged with several Petabytes of storage capacity and dynamically assignable computing resources for hosted processing. Still, a number of important issues need to be taken into account in order to properly operate and maintain an EO data exploitation platform:

- What data is to be newly ingested from external sources?
- What data is currently available under which conditions, through which access services, for which users?
- How is input data identified and accessed within the (potentially distributed) platform for a specific processing task?
- What obsolete data is to be / has been removed and when?
- How is the exploitation platform use reported?

These questions are addressed by online EO data management. Similar to the payload data management functions within satellite mission payload data ground segments, online data management is “EO-aware”, i.e. uses specific data, services, and processing properties to manage the data on the platform. The retention time of online products may for example depend on data quality, cloud cover and geolocation. Data access and processing needs on EO data exploitation platforms are much less predictable, calling for dynamic scaling and optimizing local data flows.

However, long-term preservation and continuous systematic processing of raw data remain important tasks of EO data centers such as the German Satellite Data Archive (D-SDA) [1].

3. ONLINE DATA MANAGEMENT

3.1. Context and Layered Architecture

The layered architecture depicted in Figure 1 shows online data management in its context. Large volumes of data are collected from heterogeneous external data sources and made available to the attached processing chains. In order to establish management capabilities for these data, an *Online EO Data Management* function residing within the Management & Exploitation Layer is introduced. The User Service Layer built on top consists of various data access services providing harmonized access to the hosted data for individual users. The architecture also takes into account access to an Archiving Layer containing the long-term archive which may be located at an EO datacenter and thus needs to be decoupled from the online data management and exploitation tasks which are often performed in the cloud.

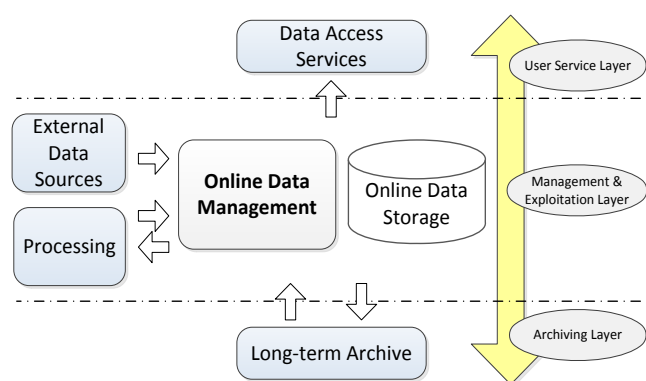


FIGURE 1: LAYERED ARCHITECTURE

3.2. Goals

The following main goals shall be addressed by an Online EO Data Management function:

- *Handling of heterogeneous data:* Heterogeneous EO data from various sources need to be ingested, managed and accessed. This also includes interoperability and data circulation between different exploitation platforms.
- *Integration of data access services:* A standardized way to access data through specific data access services is feasible. These services need to be integrated with a system providing online data management capabilities.
- *Access to the long-term archive:* The long-term archiving of data remains one of the major objectives in the EO domain. Transparent access to these backend archives is necessary if products already evicted from

the online data storage are requested by processing or data access services respectively.

- *Cloud support:* Seamless integration with cloud solutions, e.g. access to several cloud storages, resulting from the trend to move data management and processing into the cloud, taking security aspects into account.
- *Costs optimization:* Often costs are not only determined by storage space used, but also by utilized network bandwidth. Therefore, reducing copying steps by working directly on the data shall be enforced by an Online EO Data Management function whenever reasonable. Moreover, automatic migration of data to less costly storage solutions taking access statistics into account shall be supported.

3.3. Functional Components

The analysis of the presented goals leads to the identification of the following building blocks for online data management from a pure functional view:

- *Data ingestion:* Workflow-based ingestion of large volumes of data using push and pull mechanisms from heterogeneous sources. Data integrity shall be ensured on the data level as well as on the management level.
- *Data inventory:* Catalog of managed EO data to allow keeping track of the data on the platform and providing fast search capabilities.
- *Data access:* Decouple data access services by providing subscription and notification. Authorization and authentication on all kinds of data requests needs to be guaranteed.
- *Data eviction:* Systematic eviction per collection using a rule-based engine, e.g. employing a least-recently-used strategy and applying EO-specific conditions.
- *Data reload:* Reload EO data from backend archives transparently. Used primarily in reprocessing campaigns.
- *Data storage:* Hiding the details of the underlying storage system by providing a unified interface to local disk as well as several cloud storage solutions (e.g. object storages).
- *Monitoring and Reporting:* Typical cross-cutting concerns affecting every functional component. Events triggered by the components (e.g. on each workflow step) are collected and stored in a data warehouse, which then acts as the primary source for monitoring and reporting purposes. Flexible report definitions are necessary in order to support new EO missions easily.
- *Client access:* Provide means allowing clients to get transparent access to the Online EO Data Management function, supporting operations like retrieving or ingesting EO products. This includes an SDK (Software Development Kit), which can be used to access the system's interface in one's own application, as well as a feature-rich command line client.

4. ARCHITECTURAL CONCEPT

From the functional building blocks described in section 3.3, we derive an architectural concept which shall serve as the basis for a detailed design and a reference implementation of generic online EO data management services.

4.1. Data Structures

The higher-level unit managed is the *Online Product*. Online products are used to structure EO data logically and they consist of one or more objects. Each object has an associated type denoting the information it holds, e.g. DATA (original EO data file), METADATA (metadata file), BROWSE (browse image file), AUX (auxiliary data file), and others.

Objects are individually identifiable and accessible and can be in relationship to other objects. Such a relationship either expresses some kind of dependency (e.g. a processing dependency), or it is an association. For instance, a DATA object may have a dependency to an AUX object and associations to a METADATA object and a BROWSE object. The association type, which is either an aggregation or a composition, may further influence the semantics of operations on objects.

The thus induced object graph can then be used to answer questions like:

- Which objects have been used to process a DATA object of an online product?
- For which online products was a given AUX object used in processing?
- What is the transitive closure of an online product?

Such questions are important for scientists (e.g. access a product including selected input products), data managers (e.g. determine for what products a low quality auxiliary data file has been used) and systems (e.g. determining dependencies for systematic reprocessing campaigns).

Online products themselves are organized into *collections* which can be basically seen as containers for online products that share common properties.

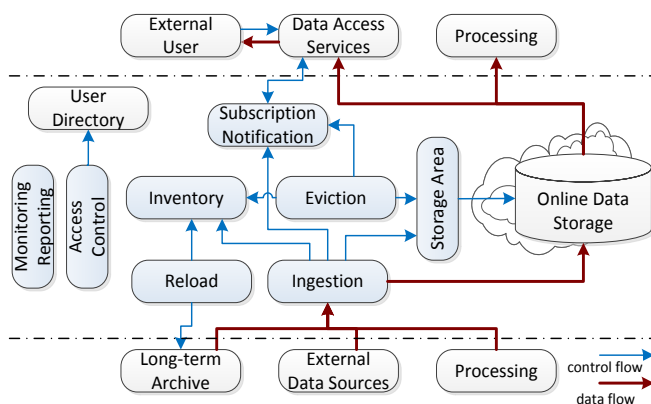


FIGURE 2: CONTROL AND DATA FLOW

4.2. Scenarios

The main online data management scenarios are shortly outlined below. They are based on the simplified control and data flow between loosely coupled services shown in Figure 2, derived from the previously introduced building blocks.

- *Ingest single product or bulk of products:* Products are divided into objects, depending on the collection they belong to. After optional processing steps, each object is ingested individually into the Online Data Storage, which is managed by Storage Areas, and the object relations are stored in the Inventory. Consistency among the Online Data Storage and the Inventory must be ensured. Load distribution using vertical scaling may be employed, but requires highly scalable data storage.
- *Access single product or bulk of products:* Transparently determine the locations of selected objects. Users or systems may then work directly on the objects, if permitted. Optionally download each object individually, and finally assemble the objects together to create the product.
- *Reload product or bulk of products:* If products are requested that have already been evicted, retrieve them from long-term archives. The LTA interface shall support priorities (e.g. for collections, users etc.) and flow control mechanisms in order to avoid LTA overloading. A successful reload scenario always triggers the ingest scenario.

5. CHALLENGES

The goals and their implications described in this paper pose several challenges, of which in our view, the following are of particular importance:

- *Bulk data handling:* Online EO data management services needs not only be able to handle heterogeneous data, but also many EO data products *at once*. The impact on the architecture still needs to be thoroughly analyzed.
- *LTA interface:* Providing access to long-term archives in order to support data reload is one of the major goals of the system. This, however, requires a standardized LTA interface, also supporting bulk data handling operations.
- *Handling of large EO products:* Optimizing bandwidth utilization is one of the key requirements for handling large EO products. Partly, this is achieved by splitting a product into objects as described in 4.1 instead of employing a single large archive file. However, this might not be sufficient, and therefore bringing the “algorithms to the data” instead of the “data to the algorithms” is a major challenge.

- *Data migration*: Although online EO data management services manage *online* data, these data may be located on a storage that is not suitable for certain data operations incurring in some processing and exploitation platform scenarios. This may be due to significant differences regarding costs of the varying storage technologies on the market, including cloud-based solutions. Therefore, data migration strategies using different storage classes shall be taken into account.

6. PROJECTS

The Sentinel-5 Precursor payload data ground segment [2] and the German Copernicus Access and Exploitation Platform project CODE-DE [3] developed by DLR both posed the needs and challenges for managing large EO data.

6.1. Sentinel-5 Precursor

The functional building blocks for S5P are similar to the ones described in this paper. S5P products are ingested with the frequency depending on near-real time or offline data and with a size of up to 2 TB/day, registered in an inventory, physically stored in a data repository of about 90 TB capacity, evicted according to specific rules and reloaded from the DLR operated S5P LTA.

However, there are several differences between the requirements for S5P and for generic online EO data management services presented in this paper:

- S5P products are only medium-sized.
- There is no exploitation platform and there are no complex data access services (only basic download via FTPS and HTTPS is provided).
- The payload data ground segment processing platform and the online data management and access system are completely separated and the data flows only in one direction, namely from the processing platform. Therefore, in order to retrieve products, the LTA is interfaced by the processing platform directly.
- The processing platform and the online data management and access system are operated on own facilities, thus no integration with cloud providers was necessary.

6.2. CODE-DE

With a daily volume of up to 8.6 TB of new data, CODE-DE has more challenging demands on the platform, currently providing one PB of online storage capacity. However, even this storage capacity is not sufficient for a permanent holding of all data within the platform. Therefore, it is necessary to introduce mechanisms as described in this paper:

- *Data ingestion*: In addition to the already described ingestion functions, the data is also registered in the CODE-DE Access Services.
- *Data inventory*: Serves as a catalog providing quick EO search capabilities and furthermore, determine if data is currently available online or needs to be reloaded.
- *Data eviction*: The employed eviction rules guarantee availability of S1, S2 and S3 data for 36 months covering Central Europe, while other coverages are available for a maximum of six months.
- *Data reload*: Reload already evicted data using a non-standardized interface to the DLR LTA.

As with S5P, the implementation of some of these functions for CODE-DE serves as precursor for generic Online EO data management services.

7. CONCLUSION

In this paper we outlined the needs for managing large EO data and approached them by introducing an Online EO Data Management function whose goals and functional components have been described. The derived generic architectural concept acts as a reference model supporting basic scenarios consisting of data ingestion, data access, data reload, and others not further described. Several challenges have then been posed which still need to be tackled in the future. Based on the reference model and on current platform developments, DLR step by step implements online EO data management services, allowing flexible and easy deployment on different infrastructures and environments. We expect these to become imperative for EO data exploitation platforms for future German missions and the growing Copernicus program.

8. REFERENCES

- [1] Kiemle, Stephan and Molch, Katrin and Schropp, Stephan and Weiland, Nicolas and Mikusch, Eberhard (2016) *Big Data Management in Earth Observation*. IEEE Geoscience and Remote Sensing Magazine (GRSM), 4 (3), pp. 51-58. Geoscience and Remote Sensing Society. DOI: 10.1109/MGRS.2016.2541306 ISSN 2168-6831
- [2] Kiemle, Stephan and Knispel, Robert and Schwinger, Maximilian and Weiland, Nicolas, *Sentinel-5 Precursor Payload Data Ground Segment*, proceedings of ESA Advances in Atmospheric Science and Applications ATMOS 2012 conference, ESA Special Publication SP-708 (CD-ROM), Bruges, Belgium, 18-22 June 2012, (2012)
- [3] Reck, Christoph and Campuzano, Gina and Dengler, Klaus and Heinen, Torsten and Winkler, Mario, *German Copernicus Data Access and Exploitation Infrastructure*, Big Data From Space BiDS'16, Tenerife, ESA, (2016)

ORGANIZING ACCESS TO COMPLEX MULTI-DIMENSIONAL DATA: AN EXAMPLE FROM THE ESA SEOM SINCOHMAP PROJECT

Alexander Jacob¹, Fernando Vicente-Guijalba², Harald Kristen¹, Armin Costa¹, Bartolomeo Ventura¹, Roberto Monsorno¹, Claudia Notarnicola¹

¹ Eurac Research, Institute for Earth Observation, Bolzano, Italy, ² Dares Technology, Barcelona, Spain

ABSTRACT

In a landscape of ever-growing access to free data, the problem of how to process all this data remains a challenge. With this article, we show how we tackle these challenges within the ESA SEOM project SInCohMap, where a critical task is to host a round robin to test different classification approaches for land cover mapping utilizing information derived from complex SAR data, in particular the evolution of multi-temporal coherence from the Copernicus Sentinel-1 satellites. Access to data and processing facilities is provided by the Eurac Research Sentinel Alpine Observatory on their computing infrastructure based on free open source technology, featuring cloud computing on OpenNebula and Kubernetes, multi-dimensional data arrays on Rasdaman and web-based python development on Jupyter.

Index Terms— Copernicus, SAR, Coherence, Data Cubes, Cloud Computing

1. INTRODUCTION

Within the last years, we have seen a huge increase in terms of freely available data due to the changes in strategy of availability of earth observation from NASA [1] and later ESA in the light of the Copernicus program [2] for research and industry. Especially the Copernicus program [3] of the European Commission and the available data e.g. from the Sentinel Series of satellites provide a wealth of data, covering various parts of the electromagnetic spectrum and are fit for a host of different applications. We have now with the Sentinel 1 and 2 constellations two operational Satellites that provide high quality SAR and optical multi-spectral data in regular intervals, allowing the inclusion of time into remote sensing based analysis on a regular basis. This brings great opportunities for both industry and research, but also comes with some challenges as how to access, utilize and analyze this data. Classic remote sensing software has limited capabilities dealing with large quantities of data and especially the time component; storage requirements are huge and in order to provide results in a timely manner a sizeable and scalable processing infrastructure is required.

At the Eurac Research Sentinel Alpine Observatory [4] we are involved in a number of regional and international research projects in the domain of environmental monitoring of the Alpine and other mountain environments. Remote

sensing data is an important part of these activities and we aim at providing easy access to our researchers and project partners.

One of these project is SInCohMap [5] (Exploitation of Sentinel-1 Interferometric [6][7] Coherence [8][9] for Land Cover and Vegetation Mapping) funded by the European Space Agency. The project, led by DARES Technologies, has great need of satellite data access and processing capabilities in order to conduct the research and development needed to exploit interferometric coherence for land cover mapping. To fulfill those needs, the data is hosted and processed within the infrastructure of the Sentinel Alpine Observatory managed by the Eurac Institute for Earth Observation. The scope of the present work is to demonstrate how to render this complex multi-dimensional big data easily accessible within the project consortium and beyond as we are going to host a round robin on land cover classification with interferometric coherence data. The round robin will be open to a wider audience from fall 2017. The overall project and research goals are explained in section 2, the planned round robin will be described in section 3, the infrastructure and computing environment will be treated in section 4.

2. SINCOHMAP PROJECT

The main objective of this project is to develop, analyze and validate novel methodologies for land cover and vegetation mapping by using time series of Sentinel-1 data and in

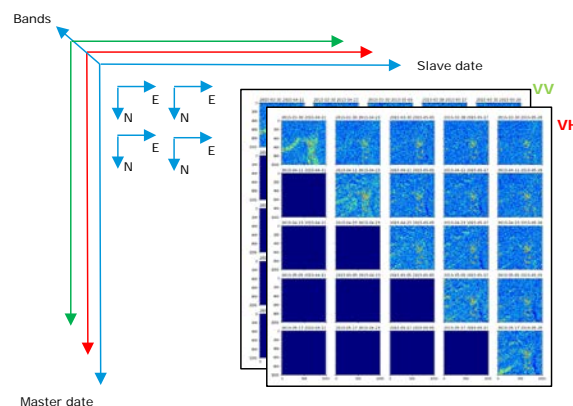


Fig. 1 Schematic view of data organization

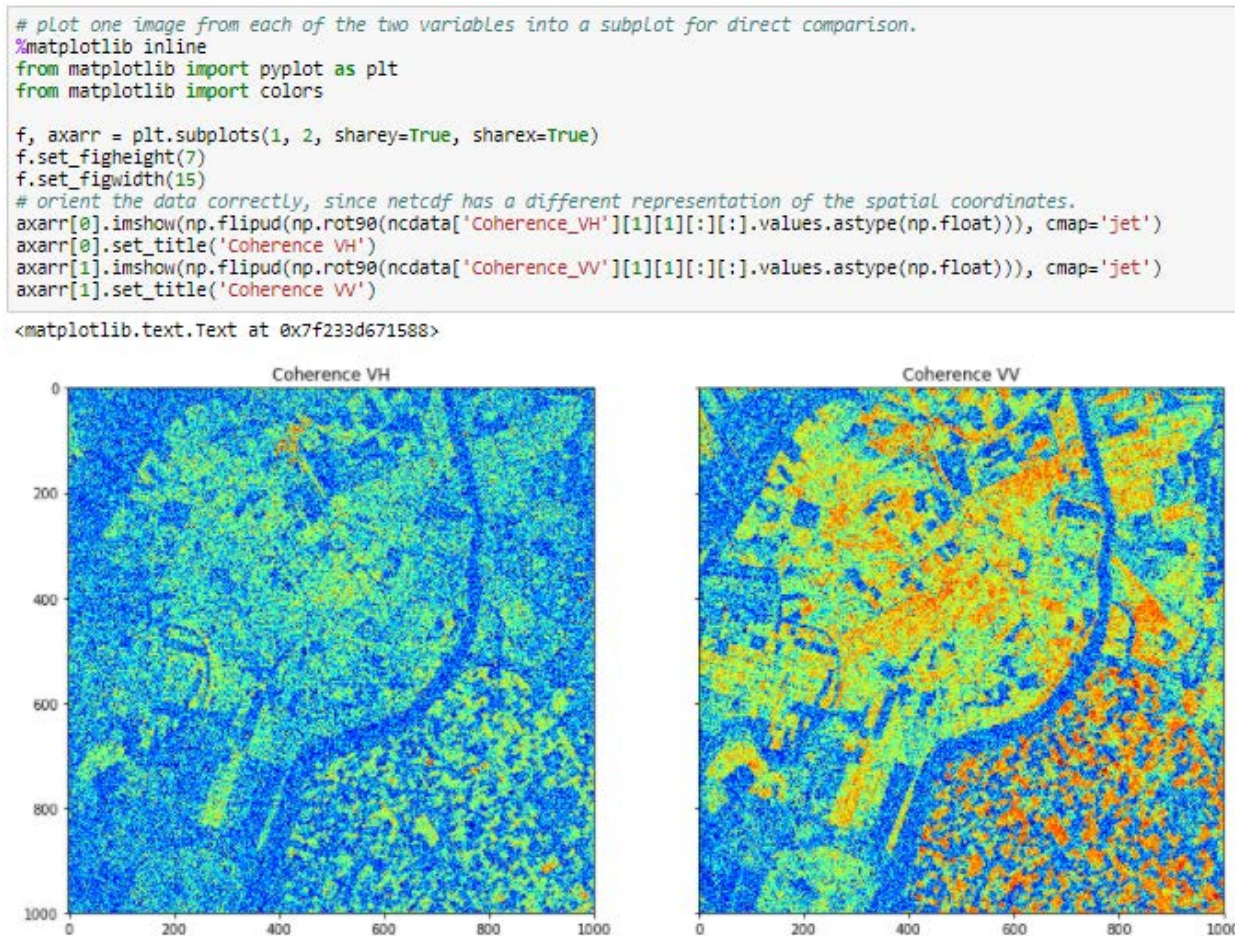


Fig. 2 Round robin tutorial sample for data access & plotting

particular by exploiting the temporal evolution of the interferometric coherence.

Since the Sentinel-1 constellation is operational, we are now in the fortunate position of having 6-day observation repeat intervals and hence can compute interferometric measures on this time scale. Additionally, this data is available in two different polarizations, increasing the possible feature space that can be derived further. We are looking at a 5-dimensional input data domain (see Fig. 1 for schematic view of the data representation) covering 2 temporal (observation times of master and slave), 2 spatial (north and east in geocoding or azimuth and range in SAR geometry) and 1 feature space (e.g. VV & VH of coherence) axis.

Further, the project aims at quantifying the impact and possible benefit of using Sentinel-1 InSAR (Interferometric Synthetic Aperture Radar) data relative to traditional land cover and vegetation mapping using optical data (especially Sentinel-2) and traditional intensity-based SAR (Synthetic Aperture Radar) approaches.

The main classes sought after are: Forests, Agricultural areas (e.g. Crops), Artificial surfaces (e.g. Urban), Water Bodies,

Scrub and Herbaceous Vegetation, Open or bare land with little to no vegetation and Wetlands.

Four different reference test areas within Europe, Spain, Italy, Poland and Finland, have been selected to cover a large variety of climate zones and land cover types with very accurate ground truth data for performing quantitative assessment and validation.

3. ROUND ROBIN

In order to scientifically evaluate the performance of different methodologies for land cover and vegetation mapping, a round robin is organized. Participants which include project partners as well as external researchers got access to pre-processed datasets over the three study areas together with some relevant training data for classification purposes. The reference data has been prepared following the well-known Corine nomenclature [10] and is available in a coarse thematic resolution of the 5 first tier classes as well as more detailed classification scheme, going down to the 3rd tier.

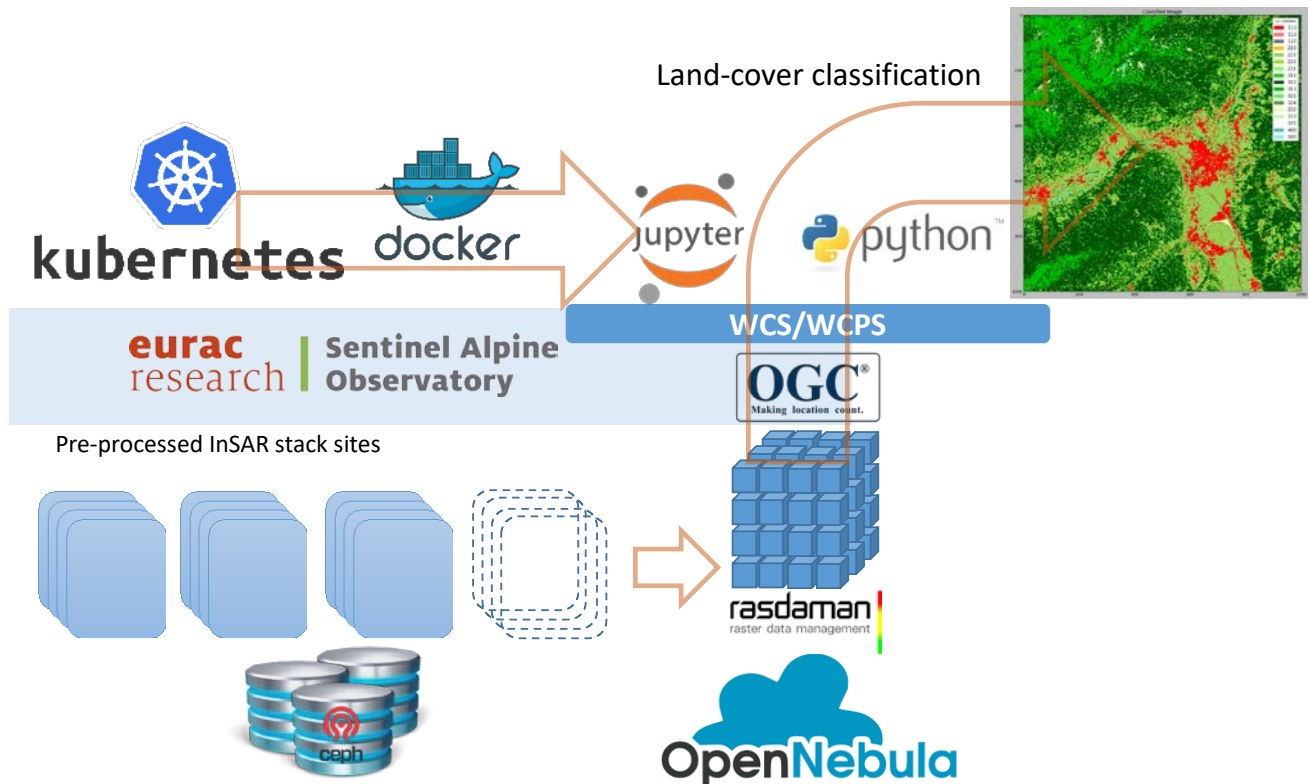


Fig. 3 Round robin technology schematic workflow

The datasets provided consists of interferometric coherence in two polarizations (VH & VV) available in both geographical projected (UTM) and in slant-range geometry. Further participants got access to processing facilities via a private cloud platform hosted at the Eurac Research Sentinel Alpine Observatory. The kickoff for this round robin was in October 2017 and it will stay active for about four month until the end of January 2018.

Tutorials with exemplary Jupyter notebooks have already been created (<https://gitlab.inf.unibz.it/SInCohMap/RoundRobinTutorials> and Fig. 2 for an example) to show how to query, access and plot the data from the Rasdaman servers and perform typical remote sensing tasks such as classifying data with machine learning techniques like Random Forest or Support Vector machines. An overview of the technology workflow can be seen in Fig. 3 and is further described in the next section.

4. PROCESSING ENVIRONMENT

The SInCohMap project infrastructure is currently hosted on the Eurac Research Sentinel Alpine Observatory infrastructure.

The Sentinel Alpine Observatory is built on a powerful processing and storage cluster for accessing and exploiting the potential of Sentinel data in the Alpine region. Having both the storage and processing facilities closely connected guaranties efficient development of satellite-data based

products. The underlying infrastructure is based on CEPH cluster storage [11] of currently 1.4 peta bytes of raw storage and a computing cluster of about 300 CPU cores and 3 TB of RAM. The storage and processing facilities are linked via 2x40Gbps network interface. The computing cluster is driving a cloud environment installed with the OpenNebula [12] cloud management framework to host a virtual private cloud infrastructure. The ceph storage, a software defined object storage, is based on commodity hardware and hence provides a relatively cheap way of organizing cluster storage with parallel access capabilities.

During the run of the SInCohMap project, additional data has been added to this infrastructure in order to cover the four principal study areas of the project. Having all data required for processing in one place and in very close coupling with the processing resources guarantees efficient development and convenient access.

Data access is provided primarily through OGC web services. Most prominently the WCS (Web Coverage Service) and the WCPS (Web Coverage Processing Service), which allow access directly to the data via http requests. To this end, the Rasdaman [13-15] implementation was chosen. It was selected as a compromise between easiness to use and performance benefits over other solutions that we found in some testing that we performed prior to settle on this solution. More details about this can be found in the master thesis of Harald Kristen, once it is published. Being able to flexible

define and serve multi-dimensional data, is the biggest perk here. As mentioned in section 2 the coherence data is represented in 5 dimensions, which in this infrastructure are all queryable and hence allow for development of classification algorithms taking the time dimension into account easily and efficiently.

We are providing access to the data and processing for users with varying degrees of IT-knowledge. The easiest access is granted developing directly in a web browser environment based on python called Jupyter [16], where execution of the code is performed on dedicated servers on the infrastructure. As said earlier tutorials for this are already existing to facilitate the access further.

After a successful implementation of the jupyter processing environment, this was translated into a docker container. Finally, with the use of a processing cluster utilizing kubernetes [17], this work environment was rendered scalable using jupyterhub for launching and hosting enough instances of the Jupyter servers for processing of the data. Each launched docker container is tightly coupled with a persistent storage access on a per user basis, to enable workflows to persist over severable life cycles of a container. Access is granted on request bases via http on <https://sincohmap-hub.eurac.edu/>.

5. CONCLUSION

This work showed an example of how to render access to complex large volumes of data possible and allow exploitation of this data in a typical research environment consisting of different projects and partners. By hosting all data and processing in one location data duplication is avoided and consistent access is granted to all interested parties. Collaborative development and research is fostered. Technologies for data access and processing developed in this project can and will be applied for other projects. We are now very excited for the scientific outcome of this exercise, which will be further, evaluated after the conclusion of the round robin in the first quarter of 2018.

6. REFERENCES

- [1] Michael A. Wulder, Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, Curtis E. Woodcock, Opening the archive: How free data has enabled the science and monitoring promise of Landsat, In Remote Sensing of Environment, Volume 122, 2012, Pages 2-10
- [2] Aschbacher J. (2017) ESA's Earth Observation Strategy and Copernicus. In: Onoda M., Young O. (eds) Satellite Earth Observations and Their Impact on Society and Policy. Springer, Singapore
- [3] <http://www.copernicus.eu>, last accessed 2017-07-31.
- [4] <http://sao.eurac.edu/>, last accessed 2017-07-31

- [5] <http://sincohmap.org/>, last accessed 2017-07-31

[6] Massonnet, D. y Rabaute, T. (1993). Radar interferometry: limits and potential. Geoscience and Remote Sensing, IEEE Transactions on, 31(2):455–464.

[7] Rosen, P. A., Hensley, S., Joughin, I. R., Li, F. K., Madsen, S. N., Rodriguez, E., y Goldstein, R. M. (2000). Synthetic aperture radar interferometry. Proceedings of the IEEE, 88(3):333–382.

[8] Lee, J.-S., Hoppel, K. W., Mango, S. A., y Miller, A. R. (1994). Intensity and phase statistics of multilook polarimetric and interferometric SAR imagery. Geoscience and Remote Sensing, IEEE Transactions on, 32(5):1017–1028.

[9] Tough, R., Blacknell, D., y Quegan, S. (1995). A statistical description of polarimetric and interferometric synthetic aperture radar data. En Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, volumen 449, páginas 567–589. The Royal S

[10] Bossard, M., et al. "The revised and supplemented Corine land cover nomenclature." European Environment Agency, Copenhagen (2000).

- [11] <http://ceph.com>, last accessed 2017-07-31

- [12] <http://opennebula.org/>, last accessed 2017-07-31

[13] Peter Baumann, Paula Furtado, Roland Ritsch, and Norbert Widmann, "The rasmaman approach to multidimensional database management," in Proceedings of the 1997 ACM symposium on Applied computing. ACM, 1997, pp. 166–173.

[14] P. Baumann: Array Databases and Raster Data Management. In: T. Özsu, L. Liu (eds.): Encyclopedia of Database Systems, Springer, Heidelberg, 2009

[15] D. Misev, P. Baumann: A Database Language More Suitable for the Earth System Sciences. G. Lohmann et al (eds.): Towards an Interdisciplinary Approach in Earth System Science. Springer 2015, doi:10.1007/978-3-319-13865-7

- [16] <http://jupyter.org/>, last accessed 2017-07-31

- [17] <https://kubernetes.io/>, last accessed 2017-07-31

LARGE SPATIAL SCALE GROUND DISPLACEMENT MAPPING THROUGH THE P-SBAS PROCESSING OF SENTINEL-1 DATA ON A CLOUD COMPUTING ENVIRONMENT

Claudio De Luca¹, Manuela Bonano^{1,2}, Francesco Casu¹, Riccardo Lanari¹, Michele Manunta¹, Mariarosaria Manzo¹, Ivana Zinno¹

1. IREA-CNR, Italy

2. IMAA-CNR, Italy

ABSTRACT

In the last decades Earth Observation (EO) from space has very fast evolved through the development of remote sensing data-acquisition systems, contributing to the creation of a Big EO Data scenario. In this work, we present a Cloud Computing solution for the advanced interferometric (DInSAR) processing chain based on the Parallel SBAS (P-SBAS) approach, aimed at processing Sentinel-1 (S1) Interferometric Wide Swath (IWS) data for the generation of large spatial scale deformation maps and corresponding displacement time series in an efficient, automatic and systematic way.

The presented approach has been used to perform a national-scale DInSAR analysis over Italy, involving the processing of more than 3000 S1 IWS images acquired from both ascending and descending orbits. Details on the cloud infrastructure and processing times will be presented.

Index Terms— DInSAR, P-SBAS, Cloud, Amazon

1. INTRODUCTION

The current EO scenario is characterized by a huge availability of Synthetic Aperture Radar (SAR) data that have been acquired during the last 25 years by past and present sensors. These data include the long-term ESA archives collected by the completed ERS-1, ERS-2 and Envisat missions. Moreover, we have disposable the data provided by the currently operational X-band generation SAR sensors, such as COSMO-SkyMed (CSK) and TerraSAR-X (TSX).

In this context, a crucial role is played by the recently launched Sentinel-1 (S1) constellation that, with its global acquisition policy, has flooded the scientific community with a huge amount of data acquired over large part of the Earth on a regular basis (down to 6-days with both Sentinel-1A and 1B passes). Moreover, the Sentinel-1 huge data archives are already fully available for the scientific community thanks to the “free and open access” distribution data policy.

Among several SAR data exploitation methodologies, we focus our attention on Differential SAR Interferometry (DInSAR), which is a well-established microwave remote

sensing technique that allows estimating the ground deformations with centimeter to millimeter accuracy [1]. Over time, DInSAR has moved from the analysis of single deformation episodes towards the study of the temporal evolution of the detected displacements, especially thanks to the availability of the above-mentioned large SAR data archives. A very well known DInSAR algorithm is the one referred to as Small BAseline Subset (SBAS) [2], which is able to generate mean deformation velocity maps and displacement time series from multi-temporal SAR datasets; it is, besides, capable to perform analyses at different spatial scales and with multi-sensor data.

Recently, an advanced parallel computing algorithmic solution, referred to as P-SBAS, that encompasses diverse parallelization strategies, both multi-nodes and multi-cores, with good scalable performances, has been developed [3]. Starting from this interferometric processing chain implementation, we developed a new framework that is specific for Sentinel-1 data (P-SBAS S1).

Such a DInSAR processing chain is capable to automatically ingest Sentinel-1 SLC SAR images and to carry out several processing steps, such as SAR image coregistration, interferogram generation, interferometric phase unwrapping, in order to finally compute mean deformation velocity maps and corresponding time series in an efficient and systematic way. In particular, we took advantage of the intrinsic structure of the S1 SAR data (acquired with the innovative TOPS mode) made of bursts considered as separate acquisitions, for deploying the coarse granularity parallelization strategy for some major steps of the P-SBAS processing chain..

Consequently, the use of Cloud Computing (CC) environment represents a promising solution to overcome the big issue relevant to the massive processing that will inevitably follow the expected huge SAR data flow provided by S1 constellation. Moreover, in-house High Performance Computing infrastructures can be very expensive in terms of procurement, maintenance, and upgrading. In addition, CC can be extremely helpful for both resource optimization and performance improvements, implying a further push towards the use of such a technology also in scientific applications.

In this work we present a Cloud Computing solution for the Advanced DInSAR processing chain for S1 Interferometric Wide Swath (IWS) data for the generation of large spatial scale deformation time series in efficient, automatic and systematic way. Different parallel strategies have been *ad hoc* designed for each processing step of the P-SBAS S1 chain, encompassing both multi-core and multi-node programming techniques, in order to maximize the computational efficiency achieved within a Cloud Computing environment and cut down the relevant processing times.

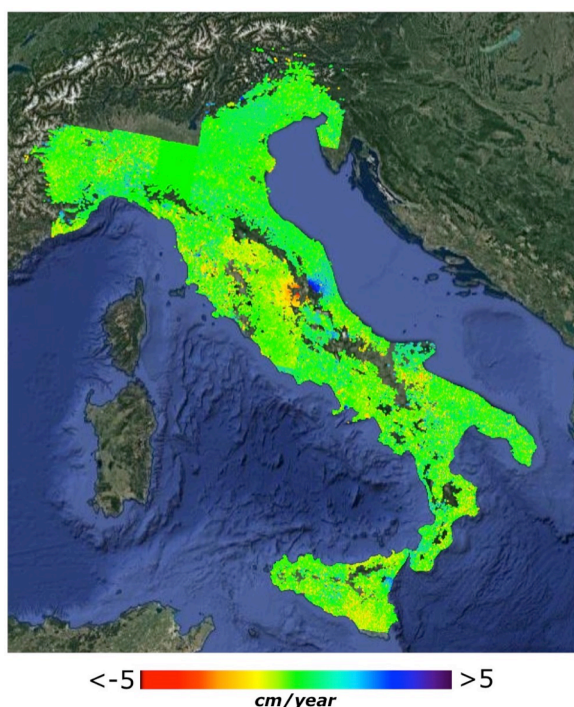


Figure 1: LOS mean deformation velocity map over the Italian territory. The results have been obtained by exploiting S1 data archives over descending orbits from October 2014 to April 2017.

2. THE P-SBAS S1 NATIONAL SCALE ANALYSIS

The presented approach was used to perform a national-scale DInSAR analysis over Italy. In particular, we exploited the SAR data acquired by the S1 constellation over descending orbits (track number: 66, 168, 95, 22, 124, 51) and ascending ones (track number: 88, 15, 117, 44, 146).

For the P-SBAS analysis we processed approximately 3.000 S1 IWS images acquired from October 2014 to April 2017, and generated about 6.500 differential interferograms, with a resolution of about 80 meters square. Starting from these interferograms, the displacement time series and the corresponding mean deformation velocity maps have been achieved. In Figure 1, the overall Line of Sight (LOS) mean deformation velocity map relevant to the processing of S1 SAR data acquired over descending orbit is shown. In

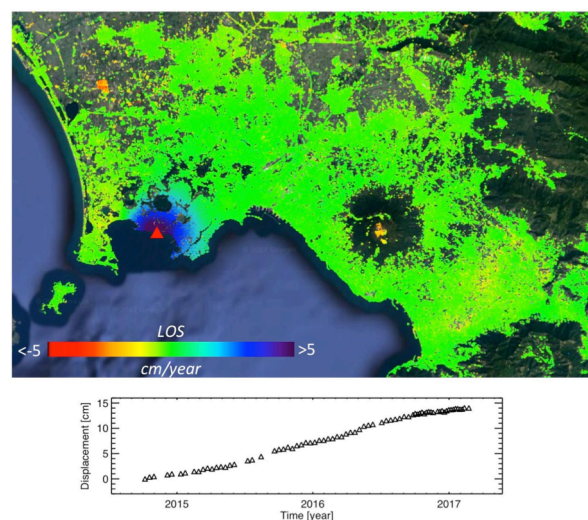


Figure 2: Zoomed view of the results shown in Figure 1. On the top, LOS mean deformation velocity map acquired over S1 descending orbits relevant to the Napoli Bay area. On the bottom, plot of the displacement time series relevant to a point of maximum deformation (red triangle).

Figure 2, instead, an inset relevant to the deformation pattern affecting the Napoli Bay area and the displacement time series of a pixel located at the maximum deformation area of the Campi Flegrei caldera are also highlighted. To cover the overall Italian territory on both ascending and descending orbit, we processed 42 independent S1 data-sets separately. The processing was performed by using the Amazon Elastic Compute Cloud (EC2) of the Amazon Web Service (AWS); in particular, a three years reserved instance has been exploited. Such an instance is equipped with 64 CPU, 500 GB of memory and 8 SSD disks with a capacity of 2 TB for each of them and a bandwidth of 20Gb/s. With these hardware characteristics we estimated to exploit about 18 weeks to process the overall Italian territory on ascending and descending orbit, thus implying 3 possible updates per years concerning the time series generation. This experiment is pioneer for the built up of an operational service for the national scale ground deformation mapping.

3. REFERENCES

- [1] D. Massonnet and K. L. Feigl, "Radar Interferometry and its application to changes in the Earth's surface," *Rev. of Geophys.*, vol. 36, pp. 441–500, 1998.
- [2] P. Berardino et al., "A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2375–2383, Nov. 2002.
- [3] F. Casu et al., "SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 8, pp. 3285–3296, Aug. 2014.
- [4] I. Zinno et al., "A Cloud Computing Solution for the Efficient Implementation of the P-SBAS DInSAR Approach," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.10, no.3, pp.802-817, Mar. 2017

BIG DATA FROM ESA EARTHNET THIRD PARTY MISSION PROGRAMME: OPPORTUNITIES AND FUTURE EVOLUTION

*Giuseppe Ottavianelli¹, Mirko Albani¹, Roberto Biasutti¹, Bianca Hoersch¹, Herve Jeanjean¹,
Henri Laur¹, Bruno Schmitt².*

1) European Space Agency (ESA/ESRIN), 2) SERCO (Italia)

ABSTRACT

The paper describes the ESA Earthnet Third Party Mission (TPM) activities and its relation to Big Data from both the data management technology and exploitation perspective. It first presents the background of Earthnet and it then describes the remarkable application opportunities that the TPM products offer to Earth Observation (EO) data users. Data is available from either operational or heritage missions. The paper further discusses the future challenges of the Earthnet TPM programme. Innovative solutions from both a technical and governance perspective will be implemented to respond to: the rapid evolution of EO-based services, the increasing data volumes and diversity, the new Information Technology developments, and the capacity optimization task of the TPM operators. These solutions allow the sustainable and progressive growth of the EO end-to-end community.

Index Terms— TPM, Earthnet, EO

1. INTRODUCTION

Since 1977, the Earthnet Programme prepares, supports and complements the ESA EO Missions by providing coherent access to non-ESA TPMs via coherent standardised interfaces. While some 20 years ago the ‘Third Party’ mechanism was the only way for many European users to get access to non-European data, the rationale today has evolved:

- TPMs complement ESA missions and/or are used jointly to cross calibrate, to validate and to prepare future ESA missions
- European users are, in their majority, today accessing data from ESA, European national and non-European Earth Observation satellites data to satisfy their research/application-driven and not mission-driven needs.
- Many more Earth Observation missions are operated by more varied organizations than ever before, all providing valuable Earth Observation data through different access mechanisms and formats. Over the last 20 years, huge archives of Third Party Earth Observation data have been created, some for global, some for regional coverages, in various processing levels and formats. Accordingly, European users’ needs for standardization of product

generation from TP missions, both for historical data and operational data, have increased.

- The user requirement for simplified data access to a wider and increasing range of Earth Observation data sources, calls for an increased sharing and interoperability of the decentralized European Earth Observation data ground segment infrastructure of national facilities, capitalizing on the investment by Member States and the Earthnet Programme.

2. ESA EARTHNET PROGRAMME

Earthnet has been and is the cornerstone for international cooperation with other Space Agencies, or private mission operators and data providers worldwide. Earthnet contributes through the TPM scheme and through representation in international bodies and boards since more than 35 years and assures those international co-operation elements which require a long term sustainability in funding and are linked to international agreements approved by Council.

Earthnet ensures coherent and non-discriminatory access for all ESA Member States to TPMs EO data, establishing therefore a solidarity component among member states in ‘Observation of the Earth’ allowing Member States without own EO missions to access data from other missions of European and international origin.

Furthermore Earthnet defines ground segment standardisation, and ensures ESA’s presence in organisations, committees (e.g. UN, GEO, CEOS) and in initiatives for promoting the international use of Earth Observation data (e.g. in Africa, China). These activities ensure complementarity and cooperation of optional ESA programmes with national missions and programmes and the international cooperation. This mandatory activity, funded out of ESA’s Science, Research and Development, has triggered investments in many ESA optional programmes and supported Member States national investments in Earth Observation.

3. BIG DATA FROM THIRD PARTY MISSIONS

Third Party Mission data can be accessed through the ESA portal [1].

The total cumulative number of scientific projects using TPM data since 2008 has reached almost 5300 with an increase of ~900 registered TPM users only in the last year.

The projects are in a wide variety of applications domains, from land applications, to water and use in atmospheric science projects. An astonishing wealth of information is freely accessible to data users.

Today Earthnet provides:

- access to data of more than 25 TPMs either for preparation or continuity of European Missions or for complementary/synergistic use with ESA's own heritage missions such as Envisat, ERS, or current missions such as Sentinel and Earth Explorers as well as national EO missions;
- for smaller missions reaching 5 years after end of the satellite operations, those missions are moved under the Long Term Data Preservation (PTDP) Programme, where Earthnet takes care for the preparation of the missions for data transfer into LTDP; in case of larger mission archives such as Landsat or ALOS, Earthnet takes care of the mission preparation for LTDP and may support the continued operations even beyond 5 years; specific systematic processing campaigns were executed for those missions under Earthnet, for easier transfer to LTDP at a later stage;
- access to TPM data is provided to all ESA Member States, scientists, application development as well as other pre-operational and non-commercial environmental projects through related international agreements, using a distributed and shared ground segment consisting of national and industrial facilities;
- the basic technology, standardisation of interfaces and their operations support for those TPMs;
- furthermore, by contributing to operate the related archives, Earthnet keeps at the disposal of European users more than 35 years of TPM data and as such provides long-term continuity for science, research and application development.

Earthnet guarantees the continued and efficient access to those missions through operations of a distributed European ground segment, composed of multi-mission ESA, national Member State and industrial infrastructure, coordinated through ESA. Within this decentralized ground segment set-up, ESA manages contracts for the re-use of existing National and industrial facilities for any activity of acquisition, processing, archiving and/or distribution of TPM data to the European user communities.

Earthnet has continuously integrated over the last years an increasing number of TPMs, at a constant level of funding: in the current Earthnet Phase, data from 28 missions with 56 individual satellites (historic and operational missions) have been maintained, transferred where relevant to LTDP and made available to users. 50% of those missions are of European origin, the other 50% are

international, mainly from Canada, US, Japan, India, South Korea.

3.1. Data access

Over the past years ESA has harmonised data access interfaces, and revised all access mechanisms in accordance with the revised ESA data policy, with related adaptation of legal documents (Terms and Conditions for the use of data, for both see links at [2]). All cost related to TPMs have been waived, so that the user communities are no longer invoiced for any data usage. A coordinated TPM data access has been harmonized and an entry point to access all online TPM data have been created: <https://tpm-ds.eo.esa.int/collections/>. This page provides, for each TPM data collection, links to: a) the dissemination server (e.g. On-line Advanced Dissemination System – OADS); b) the data collection description with access information. All users registered and authorized (from EarthNet Online) are redirected to this page for TPM online data Access. It creates a closed loop where the users are always redirected to the dissemination servers, and it also facilitates operational maintenance by unlinking dissemination function and access management function. Access management function (EarthNet Online) has also been simplified for users by minimizing the number of needed user interactions to be authorized and access the TPM online data.

3.2. TPM Status

The status of the various ESA TPMs funded by Earthnet is reported below.

Optical / multispectral / hyperspectral TPMs:

- User demand for Proba-1 hyperspectral data is still high, requiring an increased imaging capacity of ascending/descending orbits;
- Access for science users to archived Spot-1/7 data and new tasking products of Spot-6, -7 and Pleiades-A/-B data is in place through an agreement with Airbus Defence & Space;
- The possibility to reduce costs by discontinuing the systematic Landsat-7/8 processing and instead provide a Real Time processing to test emergency services is under assessment;
- The Oceansat-2 data provisioning to ESA users complements Sentinel-3 data. The provision of data from India IRS-1C, IRS-1D, Cartosat-1, Resourcesat-1 and Resourcesat-2, under agreement with GAF will be reassessed owing to the low user demand.

SAR TPMs:

- Through agreements with commercial data providers, Earthnet ensures science user access to X-band and C-band SAR missions, including TerraSAR-X, Cosmo-SkyMed and Radarsat-2 (and in the future PAZ/SeoSAR);

- Access to L-band SAR data is envisaged for Argentinian SAOCOM-1 mission, pending successful launch and positive outcome of discussion with ASI and CONAE;

Atmospheric TPMs:

- The support to ODIN operations continues to be shared between ESA and SNSB (given continued satellite health);
- The support the operations of the Japan GOSAT continues and the access to GOSAT-2 data will be sought following its launch.

Furthermore, the heritage TPM data (including ALOS, NOAA AVHRR archive, SeaWifs, KompSat-2) is managed by Heritage Data Programme (LTDP+). The exceptions are the heritage Landsat missions to maintain synergy with Landsat-8 mission currently in operation and the Spot-1/5 series to maintain synergy with Spot-6/7.

4. FUTURE EVOLUTION

The Earthnet TPM programme is also responding to new challenges related to the rapid evolution of EO-based services, the increasing data volumes and diversity, and the new Information Technology development.

Large scale exploitation shall allow not only the possibility to use a diverse and large volume of data but also the possibility for a large number of users to do more with EO data in order to stimulate Earth Science and a flourishing service sector. It is expected that future science and application projects will increasingly build on a broad variety of data sources, eventually leading to higher-level products that are created with the help of almost “invisible” EO data contributions.

Information technology is also rapidly evolving, presenting service models which are now becoming standards, like Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and also new upcoming models such as Information as a Service (InfoaaS). Future infrastructure with innovative cloud storage, hosted processing and exploitation platforms will be certainly pursued.

4.1. Hosted processing pilot

ESA has recently initiated a pilot project to offer users also the possibility to process TPM data in hosted processing infrastructures [3]. In this case, users do not download the original EO data, but only view and process it, and download results of their processing. In such a model, a number of challenges need to be tackled, as licence conditions of EO Data-as-a-Service (DaaS), allocation of processing capabilities to users, connections of the processing infrastructure to TPM data providers for ordering and archive access, etc. The purpose of this project is to gain experience with regard to those challenges and to test user acceptance of such a model. Scope of the project is to

provide to scientific users access to EO data in a hosted processing infrastructure allowing them to process the data and download the results.

4.2. Long term innovation

While most ESA and EU mission (Sentinels) operations concepts are based on a systematic acquisition scheme, many TPM operators do not fully exploit the maximum daily imaging capacity their missions may have. This is due to the fact that related mission planning, image data download, processing and archiving is costly, thus many mission operators only task an acquisition if required by a customer. This represents a key future improvement opportunity for Earthnet in order to coordinate and encourage the use of the potential spare TPM capacity.

Innovative solutions from both a technical and governance perspective will be tested to respond to these new needs and challenges. Such solutions will be based on principles of federation, collaboration, sharing, networked governance and affiliation, and they will ultimately allow the sustainable and progressive growth of the EO end-to-end community.

5. CONCLUSIONS

The ESA Earthnet TPM programme represents a unique Big Data source for Earth Observation (EO) exploitation users. Innovative solutions from both a technical and governance perspective will be implemented to respond to new challenges and allow the sustainable and progressive growth of the EO end-to-end community.

6. REFERENCES

- [1] ESA Third Party Mission
<https://earth.esa.int/web/guest/missions/3rd-party-missions/overview>
- [2] ESA Earth Online
Terms and Conditions for the use of data, for both see links at <http://earth.esa.int>
- [3] EOhopS
<https://eohops.cloudeo.store/>

ON THE SHOULDERS OF GIANTS: PROTOTYPING THE HERO VIRTUAL RESEARCH ENVIRONMENT FOR DATA VALORISATION OF HERITAGE MISSIONS

Mirko Albani¹, Joost van Bemmelen¹, Giancarlo Rivolta²

¹European Space Agency, ²Progressive Systems Srl

ABSTRACT

This paper addresses the state of the art, rationale and motivation for the prototyping of the virtual research environment for the valorization of Earth observation heritage missions. The prototype has been defined initially for Earth Observation missions owned by the European Space Agency (ESA) starting from the long standing work performed at ESA on the topic and the experience gained by a wealth of related initiatives, research and user support services and programmes.

Index Terms— Data Valorization, Long Term Data Preservation, Virtual Research Environments, Heritage Mission, Earth Observation

1. INTRODUCTION

Virtual Research Environments are popular since more than a decade [1]. Several initiatives have been established both in the framework of EU R&D programs, as well as at national and international industry level in order to establish and promote the community building in various research domains as well as to support the scientific collaboration and the sharing of experience, data and scientific results. From the point of view of the collaboration among scientists or researchers the virtual research environments have been first conceived as “virtual organizations” in the context of the Grid computing infrastructure [2]. Within the Earth Observation Programme, there is a long tradition of work on various related topics, from the Grid computing paradigm [3], [4], through the participation to several EU funded R&D projects like e.g. D4science (<https://www.d4science.org/>), to the definition of a model for research and service support [5], [6] and of the “open science program” [7]. The latter identifies the scientific process: from Conceptualization, to Data Gathering, Analysis, Publication and Review, and an overarching interconnection of trends and changes or opportunities (like open access publications, alternative reputation systems, citizen science, open data access, online courses, etc.) which allow new actors to be involved, like the citizens or accelerate scientific processes diminishing barriers and empowering collaboration.

Other related initiatives go under the name of exploitation platforms, of which the Proba-V Mission Exploitation

Platform (<https://proba-v-mep.esa.int/>) [8] which encompasses both geospatial analytics and time series analysis functionality was the first to go in operations in January 2016, or the Geohazard Exploitation Platform [9], which focuses on the community contribution. In this short overview of virtual research environment related activities we should mention as well the Sentinel Application Platform (SNAP), a.k.a. the Sentinel Toolbox and the so called RSS Cloud Toolbox provided by the ESA RSS Service to registered users (see Figure 1).

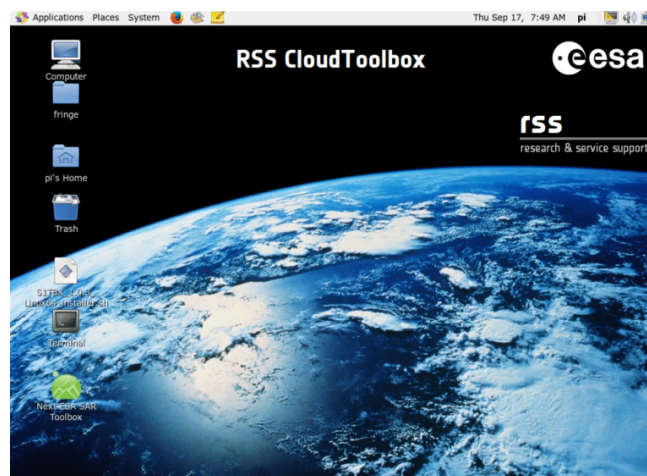


Figure 1. RSS Cloud Toolbox can support the specific needs of the Earth Observation community providing CloudToolboxes that are equipped with the latest ESA and third-party tools for EO data processing (e.g. SNAP, BEAM and PolSARPro).

The RSS CloudToolbox service [11], provides customised virtual machines (VMs) made available on a (European) Cloud Provider Infrastructure. Thanks to high-speed network connections, these cloud VMs have fast access to the ESA heritage mission data as well as for the long time series to more recent Copernicus data. They are provided with pre-installed software supporting EO data exploitation, like SNAP [10], PolSARPro, GDAL, etc. Additional software like Jupiter Notebook [12], can be installed on request.

The main objective of the RSS CloudToolbox service is to provide EO data users with proper resource flexibility, accessible via their own devices (PC, laptop, tablet) from anywhere in the world, to easily perform

their own processing. High-speed data access and powerful hardware resources from the cloud are combined to enable scientists, service providers or students to test and run own algorithms on many EO datasets available on-line.

The type of activities supported by the RSS CloudToolbox service are:

- University courses
- Workshop courses
- Research institutions
- SME developing new services
- Technology projects
- Exploitation Platform
- Individual researchers

As examples of such activities supported by RSS we can mention among many others the provisioning of customised toolboxes to TU Delft students attending Interferometry courses, or to La Sapienza University Data Science students attending lessons on Earth Observation data exploitation. Furthermore, it is relevant to mention as well the support to researchers working on various different research topics, ranging from evaluation of the potential of optical time series for improving Land Cover classification (WUR) [13] to algorithm development for generating decorrelation products from pairs of radar images (BRGM).

For the purpose of this paper the term Virtual Research Environment (VRE) is used [1] with a comprehensive scope, representing a concept overarching most of the environments cited above and an environment with the following distinguishing features: (i) it is web-based; (ii) it is tailored to serve the needs of the EO community (actually several sub-communities scientists, service providers, universities are explicitly targeted) ; (iii) it is expected to provide the commodities needed to accomplish the community's goal(s); (iv) it is open and flexible with respect to the overall service offering and lifetime. The actual algorithms and processing are entirely owned and controlled by the users, however, differently from [1], it offers a very open and loosely controlled forum to its participants, but the level of online collaboration allowed is very low and limited to the one offered by the supported videoconferencing tools: Skype and WebEx.

2. STANDING ON THE SHOULDERS OF GIANTS

The prototype virtual research environment defined to support the valorization of the heritage missions here discussed was named HERO: Heritage Eo data Research environment. Its most important characteristic is to foster user's standing on the shoulder of giants, offering (i) a set of pre-defined services which – designed by leading scientists – offer the possibility to test, play and possibly fine-tune already developed sophisticated algorithms. Prototypical of

this approach is the SBAS interferometric processing [6] offered on 10 years of ENVISAT data on the entire world; (ii) HERO Cloud Toolboxes preloaded with stacks of ENVISAT and Landsat products, where users can freely manipulate and process data, (iii) a (virtual) catalogue and access to most EO data available in ESA and in the world, (iv) an own forum to interact with the support team and last but not least (v) the possibility to create “own” long time series joining together heritage mission data with current Sentinel data.

The HERO prototype is currently limited to support the valorization of ESA heritage (and third party) missions, but is open to encompass national heritage missions, as well as – in the context of GEO – international heritage missions and tools.

3. REFERENCES

- [1] L. Candela, D. Castelli, P. Pagano, “Virtual Research Environments: An Overview and a Research Agenda,” *Data Science Journal*, Volume 12, 10 August 2013.
- [2] I. Foster, C.O. Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 1998.
- [3] L. Fusco, P. Goncalves, F. Brito, R. Cossu, C. Retscher, “A new Grid-based system to assist user in ASAR handling and analysis”, European Geoscience Union General Assembly, Vienna, 02, 07 April 2006.
- [4] J. Farres, E. Mathot, S. Pinto: G-POD: A Collaborative Environment for Earth Observation at the European Space Agency, Proceedings of the ESA Living Planet Symposium, 28 June- 2 July 2010, Bergen, Norway, Special Publication SP-686 on CD-ROM, ESA Publications Division, European Space Agency, Noordwijk, The Netherlands, 2010.
- [5] P.G. Marchetti, G. Rivolta, S. D'Elia, J. Farres, G. Mason and N. Gobron “A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support”, *IEEE Geoscience and Remote Sensing*(162): 10-18, 2012
- [6] C. De Luca, R. Cuccu, S. Elefante, I. Zinno, M. Manunta, G. Rivolta, V. Casola, R. Lanari, F. Casu, “unsupervised on-demand Web Service for DInSAR processing: the P-SBAS implementation within the ESA G-POD Environment, Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International, Issue Date: 26-31 July 2015,
- [7] P.P Mathieu, M. Borgeaud, Y.L. Desnos, M. Rast, C. Brookmann, L. See, S. Fritz, R. Kapur, M. Mahecha and U. Benz, “The Earth Observation Open Science Program” *ieec*

Geoscience and remote sensing magazine, pp 86-93, Digital Object Identifier 10.1109/MGRS.2017.2688704, June 2017

[8] E. Goor, J. Dries, D. Daems, M. Paepen, F. Niro, P. Goryl, P. Mougnaud, A. Della Vecchia “PROBA-V Mission Exploitation Platform”, Remote Sensing, 8(7), 564; doi:10.3390/rs8070564, 2016

[9] M. Manunta et al. The contribution of the Geohazards Exploitation Platform for the GEO Supersites community, EGU General Assembly, 2016

[10] N. Fomferra et al. “SNAP Home, Developer Guide”, <https://senbox.atlassian.net/wiki/display/SNAP/SNAP+Home>, accessed on 18.05.2017

[11] R. Cuccu et al. “RSS CloudToolbox Service”, <https://wiki.services.eoportal.org/tiki-index.php?page=RSS+CloudToolbox+Service>, accessed on 26.07.2017

[12] <https://ipython.org/notebook.html> accessed 26.07.2017

[13] J. Eberenz, J. Verbesselt, M. Herold, N. Tsendbazar, G. Sabatino, G. Rivolta “Evaluating the Potential of PROBA-V Satellite Image Time Series for Improving LC Classification in Semi-Arid African Landscapes”, Remote Sensing 8 (2016)12. - ISSN 2072-4292 - 11 p. doi:10.3390/rs8120987, 2016

LONG-TERM DATA PRESERVATION DATA LIFECYCLE AND STANDARDISATION PROCESS

Mirko Albani¹, Rosemarie Leone¹, Katrin Molch³, Razvan Cosac², Iolanda Maggio², LTDP WG⁴
European Space Agency¹, Rhea Group², DLR³, International Partners⁴

ABSTRACT

Science and Earth Observation data represent today a unique and valuable asset for humankind that should be preserved without time constraints and kept accessible and exploitable by current and future generations. In Earth Science, knowledge of the past and tracking of the evolution are at the basis of our capability to effectively respond to the global changes that are putting increasing pressure on the environment, and on human society. This can only be achieved if long time series of data are properly preserved and made accessible to support international initiatives. Within ESA Member States and beyond, Earth Science data holders are increasingly coordinating data preservation efforts to ensure that the valuable data are safeguarded against loss, and kept accessible and useable for current and future generations. This task becomes increasingly challenging in view of the existing 40 years worth of Earth Science data stored in archives around the world and the massive increase of data volumes expected over the next years from e.g. the European Copernicus Sentinel missions.

Long Term Data Preservation (LTDP) aims at maintaining information discoverable and accessible in an independent and understandable way, with supporting information, which helps ensuring authenticity, over the long term.

A focal aspect of LTDP is data curation. Data curation refers to the management of data throughout its life cycle. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and allow data re-use over time. It includes all the processes that involve data management, such as pre-ingest initiatives, ingest functions, archival storage and preservation, dissemination, and provision of access for a designated community.

Index Terms— LTDP Data Lifecycle, Preservation Workflow, PDSC, GSCB, CEOS, GEO.

1. INTRODUCTION

The paper presents specific aspects, of importance during the entire Earth observation data lifecycle, with respect to evolving data volumes and application scenarios. These particular issues are introduced in the section on 'Big Data' and LTDP. The Data Stewardship Reference lifecycle section describes how the data stewardship

activities can be efficiently organised, while the following section addresses the overall preservation workflow and shows the technical steps to be taken during data curation. The paper concludes with introducing international collaboration for developing coordinated and harmonised lifecycle concepts.

2. BIG DATA AND LTDP

'Big Data' indirectly addresses long-term data preservation issues: very large data sets handling, their curation, valorisation, retrieval, manipulation and finally visualization.

One of the most relevant 'Big Data' aspects is a new way of carrying out scientific research. Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive data sets.

Following experimental, theoretical, and computational science, a 'Fourth Paradigm' is emerging in scientific research. This refers to the data management techniques and the computational systems needed to manipulate, visualize, and manage large amounts of scientific data.

The main challenge is not only the volume of data, but its diversity, e.g. in format and type. Other major challenges are data structure and 'data on the move' i.e. transferring data through networks. This latter issue is a big inhibitor to jointly using data across distributed archives. Older Science and EO data are recorded on various devices, in different formats. A huge task represents the recovery, reformatting, reprocessing of such data, as well as the transcription of various associated information, necessary to understand and use the data. Challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization. A large proportion of users are not domain experts anymore, therefore data discovery tools, documentation and support are also needed.

3. DATA STEWARDSHIP REFERENCE LIFECYCLE

Earth Science data curation and preservation should be addressed during all mission stages - from the initial mission planning, throughout the entire mission lifetime,

and during the post-mission phase. The Data Stewardship Reference Lifecycle (Figure 1) gives a high-level overview of the steps useful for implementing curation and preservation rules on mission data sets from initial conceptualisation or receipt through the iterative curation cycle.

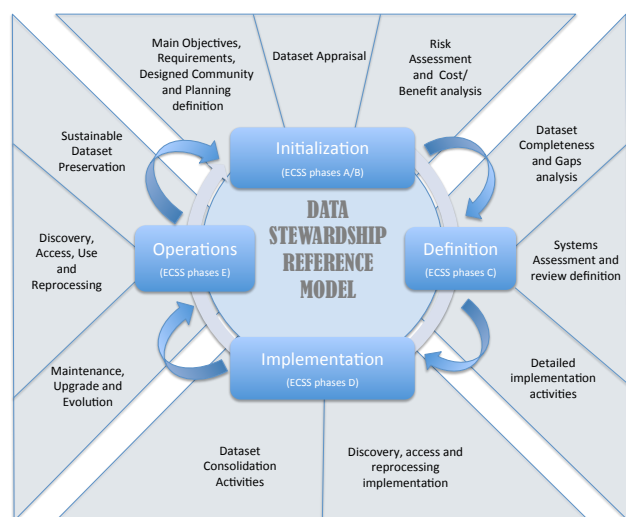


Fig. 1 LTDP Data Stewardship Reference Lifecycle

The core target of the LTDP lifecycle is the preserved data set, composed of consolidated:

- **Data records:** these include raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and descriptive information.
- **Associated knowledge:** this includes all the *processing software* used in the product generation, quality control, the product visualization and value adding tools, and *documentation* needed to make the data records understandable to the designated community. This includes among others mission operation concept, products specifications, instruments characteristics, algorithms description, Cal/Val procedures, mission/instruments performances reports, quality related information, etc. It is necessary to ensure data remain understandable and usable.

The final, consistent, consolidated, and validated “data records” are obtained by applying a consolidation process consisting of Data collection, Cleaning/pre-processing, Completeness analysis, Processing/reprocessing.

In parallel to the data records consolidation process, the data records knowledge, associated information and processing software are also collected and consolidated.

Data stewardship implements and verifies, for the relevant preserved data sets, a set of preservation and curation activities on the basis of a set of requirements defined during the initial phase of the curation exercise.

Data preservation activities focus on Earth observation data sets long-term preservation, and are tailored according to its mission specific preservation/curation requirements. They consist of all activities required to ensure the “preserved data set” bit integrity over time, its discoverability and accessibility, and to valorise its (re)-use in the long term (e.g. through metadata/catalogue improvement, processor improvement for algorithm and/or auxiliary data changes and related (re)-processing, linking and improvement of context/provenance information, quality assurance). Preservation activities for digital data record acquired from the space segment and processed on ground embrace ensuring continued data records availability, confidentiality, integrity and authenticity as legal evidence to guarantee that data records are not changed or manipulated after generation and reception over the whole continuum of data preservation (archival media technology migration, input/output format alignment, etc.), valorisation and curation activities. The usage of persistent identifier for citation is part of the agency long term data preservation best practices.

Data curation activities aim at establishing and increasing the value of “preserved data sets” over their lifecycle, at favouring their exploitation, possibly through the combination with other data records, and at extending the communities using the data sets. These include activities such as primitive features extraction, exploitation improvement, data mining, and generation/management of long time data series and collections (e.g. from the same sensor family) in support to specific applications and in cooperation with international partners.

Data stewardship activities refer to the management of an EO Data set throughout its mission life cycle phases and include preservation and curation activities. It includes all the processes that involve data management (ingestion, dissemination and provision of access for the designated community) and data set certification.

4. PRESERVATION WORKFLOW

The LTDP data stewardship reference lifecycle is also represented through the preservation workflow, which defines a recommended set of actions to be sequentially implemented for the preservation of a “data set”, with the goal of ensuring and optimizing its (re)-use in the long

term. This preservation workflow, collaboratively developed with European space data holders, ensures that Earth observation mission data sets remain accessible and useable in the long term. Applying this workflow will produce a consolidated, accessible and useable Earth observation data set – consisting of the data records and the associated knowledge – and comprehensive documentation of the preservation procedure. While best initiated during the early mission planning phases, the preservation workflow can also be applied to data sets of current and historic Earth observation missions. The preservation workflow recommended actions/steps are the following:

- EO missions/sensors data set appraisal, definition of designated community & preservation objective
- Tailoring of mission specific consolidation process
- EO missions/sensors data set PDSC tailoring and inventory table filling
- Tailored PDSC consultation with designated community
- Implementation of tailored consolidation process and collection of documentation and software
- Update of EO missions/sensors data set PDSC & inventory table
- Archive & ingestion, master inventory and catalogue population
- Dissemination & Web configuration
- Risk & cost assessment, preservation & cost planning, implementation.

5. WGISS DATA STEWARDSHIP MATURITY MATRIX WHITE PAPER

The scope of the on-going WGISS Data Stewardship Maturity Matrix definition is to measure the overall preservation lifecycle and to verify the implemented activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. It can be used to create a stewardship maturity scoreboard of dataset(s) and a roadmap for scientific data stewardship improvement; or to provide data quality and usability information to users, stakeholders, and decision makers. In the extended environment of Maturity Matrices and Models, the Maturity Matrix for “Long-Term Scientific Data Stewardship”, of Ge Peng and Jeffrey L. Privette (2015), represents a systematic assessment model for measuring the status of individual datasets. In general, it provides information on all aspects of the data records, including all activities needed to preserve and improve the information content, quality, accessibility, and usability of data and metadata. This was

used as a starting point of the WGISS Data Stewardship Maturity Matrix. In parallel, the GEO Data Management Principles Task Force was tasked with defining a common set of GEOSS Data Management Principles (DMP-IG). These principles address the need for discovery, accessibility, usability, preservation, and curation of the resources made available through GEOSS.

	DMP-1	DMP-2	DMP-3	DMP-4	DMP-5	DMP-6	DMP-7	DMP-8	DMP-9	DMP-10
Level 0
Level 1
Level 2
Level 3

Fig. 2 WGISS Data Stewardship Maturity Matrix

The content of the WGISS Data Stewardship Maturity Matrix represents the result of a combined analysis performed on the DMP-IG and a consultation at European level, with the Long Term Data Preservation Working Group. It is a self-assessment and it is applied at dataset level.

6. COOPERATION ACTIVITIES

ESA is cooperating in the LTDP domain in Earth observation with European partners through the LTDP Working Group, formed within the Ground Segment Coordination Body (GSCB), and with other international partners, through participation to various working groups and initiatives. The EO LTDP framework international context is shown in Fig. 3.



Fig. 3 EO LTDP Framework international context

The LTDP core documents have also been reviewed and approved at international level within the Committee on Earth Observation Satellites (CEOS) and the Group on Earth Observations (GEO). A review of the Preservation Workflow document is currently on going in the frame of the CCSDS Data Archive Ingestion (DAI) working group.

7. MEDIA RESCUE ACTIVITY: LESSONS LEARNED

Heritage data preservation activities include the preservation of unique data that can only be recovered from historical media. Therefore, the preservation of these media, together with the hardware that could read the media, should be ensured. During the rescue activity of JERS-1 mission media, some lessons learned were collected. Having no inventory available for the JERS-1 media at the Fucino ground station, several trips to the facility were undertaken in order to manually generate the media inventory. This was later compared against the JERS-1 data already available at ESA, which allowed to identify the missing data. However, this was not a simple task, as a large part of the media labels were either missing crucial information or this information could not be easily read, due to deterioration over time, as the storage environment was not systematically monitored.

The main lesson learned from this media rescue activity is that long-term preservation should be considered, and planned for, from the initial stages of a mission, in order to ensure that long-term data preservation policies are followed throughout the mission lifetime. Preservation of the main information on media labels and in local, digital, inventories should also be ensured, together with other Associated Knowledge. Furthermore, the original media, hardware and software should be preserved until it is certain that all unique data that could be recovered, was retrieved from the historical media. This also implies that the physical archiving storage must be located in a well-controlled environment that would prevent deterioration of the media labels or the media itself.

8. CONCLUSIONS

Data holdings are growing exponentially in Earth Science data archives worldwide. The European Copernicus program will continue to deliver Petabytes of valuable satellite-based Earth observations for many years to come.

Only a systematic approach to data preservation during the entire data lifecycle, coordinated between data holders and application communities, will ensure that these data sets will be accessible and useable to current and future generations, for monitoring long-term variations in environmental parameters as a basis for

objectively assessing and predicting effects of global change.

9. REFERENCES

- [1] CEOS, “EO Data Preservation Guidelines Best Practices”, http://ceos.org/document_management/Working_Groups/WGIS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Data%20Preservation%20Guidelines_v1.0.pdf
- [2] CEOS, “EO Preserved Data Set Content”, http://ceos.org/document_management/Working_Groups/WGIS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf
- [3] CEOS, “Long Term Preservation of Earth Observation Space Data: Preservation Workflow”, http://ceos.org/document_management/Working_Groups/WGIS/Interest_Groups/Data_Stewardship/Best_Practices/Preservation%20Workflow_v1.0.pdf
- [4] CEOS, “Associated Knowledge Preservation Best Practices”, http://ceos.org/document_management/Working_Groups/WGIS/Documents/WGISS%20Best%20Practices/CEOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf
- [5] CEOS, “Generic Earth Observation Data Set Consolidation Process”, http://ceos.org/document_management/Working_Groups/WGIS/Interest_Groups/Data_Stewardship/Best_Practices/GenericEarthObservationDataSetConsolidationProcess_v1.0.pdf
- [6] CEOS, “Long-Term Preservation of Earth Observation Space Data: Glossary of Acronyms and Terms”, http://ceos.org/document_management/Working_Groups/WGIS/Interest_Groups/Data_Stewardship/White_Papers/EO-DataStewardshipGlossary_v1.2.pdf
- [7] CEOS, “CEOS Persistent Identifier Best Practices”, http://ceos.org/document_management/Working_Groups/WGIS/Documents/WGISS%20Best%20Practices/CEOS%20Persistent%20Identifier%20Best%20Practices_v1.2.pdf
- [8] GEOSS, “Data Management Principles”, https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
- [9] NESTOR, “LONG-TERM PRESERVATION”, http://www.langzeitarchivierung.de/EN/Netzpublikationen/Langzeitarchivierung/langzeitarchivierung_node.html
- [10] BIG DATA, “Fourth Paradigm ”, http://www.astro.caltech.edu/~george/aybi199/4th_paradigm_book_complete_lr.pdf
- [11] Internet of Things (IoT), “A vision, architectural elements, and future directions” <http://www.sciencedirect.com/science/article/pii/S0167739X13000241>
- [12] RDA, “Preservation, Tools, Techniques and Policy ” <https://www.rd-alliance.org/groups/preservation-tools-%20techniques-and-policies>
- [13] Peng, Privette, Scientific Data Stewardship Maturity Matrix <http://www.slideshare.net/gepeng86/scientific-data-stewardship-maturity-matrix>
- [14] GEOSS, “Data Management Principles”, https://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf

TOWARDS A PRESERVATION CONTENT STANDARD FOR EARTH OBSERVATION DATA

Hampapuram Ramapriyan^{1,2}, Dawn Lowe², Andrew Mitchell², Kevin Murphy³

¹Science Systems and Applications, Inc., Lanham, MD, USA

²NASA Goddard Space Flight Center, Greenbelt, MD, USA

³NASA Headquarters, Washington, DC, USA

ABSTRACT

Earth observation data from a combination of spaceborne, airborne and in situ sensors have been growing rapidly over the last two decades, and have a much longer history. The observational data as well as digital products derived from them constitute a valuable global asset that must be preserved for the future generations. Currently, there are no international standards specifying the artifacts associated with the data and products that need to be preserved so that they can be understood and reused several decades in the future. However, significant efforts have been made by some national and international groups that can contribute towards such a standard. These efforts and several existing standards from the International Standards Organization (ISO) related to data, metadata and preservation are discussed briefly along with a proposal to develop an ISO standard for specifying preservation contents.

Index Terms— Earth Observation, Preservation, Standards, Data, Metadata, Provenance, Context

1. INTRODUCTION

Information from Earth observing missions (remote sensing with airborne and spaceborne instruments, and in situ measurements such as those from field campaigns) is proliferating in the world. Many agencies across the globe are generating important datasets by collecting measurements from instruments on board aircraft and spacecraft, globally and constantly. The data resulting from such measurements are a valuable resource that needs to be preserved for the benefit of future generations. These observations are the primary record of the Earth's environment and therefore are the key to understanding how conditions in the future will compare to conditions today. Earth science observational data, derived products and models are used to answer key questions of global significance. In the near-term, as long as the missions' data are being used actively for scientific research, it continues to be important to provide easy access to the data and services commensurate with current information technology. For the longer term, when the focus of the research community shifts toward new missions and observations, it is essential to preserve the previous mission data and associated information. This will enable a new user in the future to

understand how the data were used for deriving information, knowledge and policy recommendations and to “repeat the experiment” to ascertain the validity and possible limitations of conclusions reached in the past and to provide confidence in long term trends that depended on data from multiple missions.

Organizations that collect, process, and utilize Earth observation data today have a responsibility to ensure that the data and associated content continue to be preserved by them or are gathered and handed off to other organizations for preservation for the benefit of future generations. In order to ensure preservation of complete content necessary for understanding and reusing the data and derived digital products from today's missions, it is necessary to develop a specification of such preservation content. While there are existing standards that address archival and preservation in general, there are no existing international standards or specifications today to address what content should be preserved. The purpose of this paper is to outline briefly the existing standards that apply to preservation (section 2) and describe a recent effort in getting an international standard in place for specifying preservation content for Earth observation data and derived digital data products (section 3). The remaining work needed to arrive at a standard will be described in the concluding section (section 4).

2. APPLICABLE EXISTING STANDARDS

There are several standards developed by various technical committees of the International Standards Organization (ISO) that are relevant to Earth Observation (EO) data and derived digital products. A brief discussion of these is given below, summarized from the respective standard documents cited below. The reader is referred to those cited documents for complete details.

2.1. ISO 14271

Space agencies that are members of the international Consultative Committee for Space Data Systems (CCSDS) have developed recommendations titled the Reference Model for Open Archival Information System (RM-OAIS). The most recent update to the OAIS Reference Model is the publication International Standard Organization's (ISO) 14721:2012 “Space data and information transfer systems – Open archival information system (OAIS) – Reference

model”, which provides a conceptual framework for archiving [1]. This standard is designated as “CCSDS recommended practice for an OAIS reference model.” It is to be noted that the term “open” here refers to the fact that the reference model and other standards based on it are developed in open forums, and does not imply that the archive access is unrestricted. The OAIS framework addresses all functions associated with long-term preservation of information – ingest, archival storage, data management, access, and dissemination. It is applicable to all archives, but specifically those with responsibility to make information available in the long term. It opens the door for the development of more detailed standards in this area. It identifies potential areas (a “road map”) for such standards and lists several that have been developed. It does not, however, specify a design or implementation, leaving the option for breaking out the specified functionality at the implementer’s discretion.

2.2. ISO 16363

As indicated in the road map in ISO 14271, a standard has been developed by CCSDS and ISO that specifies requirements for certification of trustworthy digital repositories, based on the OAIS Reference Model. This standard, ISO 16363, is also designated by CCSDS as a recommended practice [2]. It is based on the publication by the joint task force of the Research Library Group (RLG) and the National Archives and records Administration (NARA) constituted to address digital repository certification [3]. The ISO 16363 helps auditors of digital repositories with objective criteria and metrics to assess and certify them. It also helps institutions hosting repositories to identify weaknesses and make improvements through self-assessments.

2.3. ISO 16919

ISO 16919 describes how to audit archives for compliance with the requirements [4]. Also a CCSDS recommended practice, this standard provides requirements to be met by bodies that audit and certify candidate trustworthy digital repositories. It is also helpful for digital repository staff to understand the certification process. In the context of this standard, trustworthiness of repositories means “that they can be trusted to maintain, over the long-term, the understandability and usability of digitally encoded information placed into their safekeeping.” The standard describes how a third party can inspire confidence that it has performed the certification with: impartiality, competence, responsibility, openness, confidentiality, and responsiveness to complaints.

2.4. ISO 19115-1 and 19115-2

ISO 19115 is a standard with a general title “Geographic Information – Metadata”. It has three parts.

ISO 19115-1 covers fundamentals, 19115-2 provides extensions for imagery and gridded data and 19115-3 consists of technical specifications for an XML schema implementation of metadata fundamentals. The primary purpose of ISO 19115-1 is to describe digital information that has a geographic extent [5]. However, it can be used to describe information resources that do not have a geographic extent – e.g., documents, software and repositories. Implementation of this metadata standard by resource providers will help effective and complete characterization of resources; facilitate management; enable understanding and appropriate use; help discovery, access and reuse; and assist users in deciding whether a given resource is useful to them.

ISO 19115-2 augments ISO 19115-1 with additional structure to describe more extensively the derivation of geographic imagery and gridded data [6]. It provides the structure needed to represent properties of instruments acquiring data, instrument geometry, production processes, etc. It covers metadata needed for describing derivation of “geographic information from raw data, including the properties of the measuring system, and the numerical methods and computational procedures used in the derivation.”

2.5. ISO 19157

ISO 19157:2013 standardizes components and structures of data quality measures, which are important to compare different datasets to determine which of them best fit a user’s needs [7]. It provides principles for describing data quality and concepts for handling information on data quality, and a consistent manner in which such information is determined and reported. It also provides guidelines for evaluation of quantitative quality information for geographic data.

2.6. ISO 19165

ISO 19165, a standard for “Geographic Information - Preservation of digital data and metadata” was developed during 2014-2016, and considers geographic information preservation in general [8]. It enumerates several distinguishing characteristics of geospatial data (e.g., relation to a well-defined section of the Earth, exchange by using theme-specific and sophisticated exchange formats, large data volumes, and existence of several levels-of-detail of the same dataset), and postulates that such data should be “preserved together with relevant metadata content that fully addresses these structural characteristics.” It also indicates that while its focus is on geospatial data, the principles brought forth by this standard can be applied to other types of data as well.

It defines the requirements for the long-term preservation of digital geospatial data, including “metadata, representation information, provenance, context and any

other content items that capture the knowledge that are necessary to fully understand and reuse the archived data.”

3. TOWARDS A CONTENT STANDARD

While all of the standards discussed above provide a good basis for developing a content standard for preserving Earth observation data and derived digital products, currently there is no standard that meets this need.

The standard ISO 19165 states: “In preserving data, future users need to understand what they are working with (context information) and how the data were created (provenance information). Because most Earth science data involve complex physics and mathematics, the metadata shall include sufficient documentation (or pointers thereto) that provide the derivation of the algorithms used to generate the dataset. Likewise, the metadata shall include pointers to calibration data and ancillary data that were needed to produce the dataset. The specific content items needed to preserve the full provenance and context of the data and associated metadata depend on the needs of the designated community and types of datasets (e.g., maps, remotely sensed data from satellites and airborne instruments, and physical samples). Follow-up parts to this standard may be developed detailing content items appropriate to individual disciplines.”

A New Work Item Proposal (NWIP) to develop a standard titled “Geographic information -- Preservation of digital data and metadata -- Part 2: Content specifications for Earth observation data and derived digital products” has recently been approved by the ISO Technical Committee for Geographic Information/Geomatics (TC 211). This new standard, ISO 19165-2, will provide detailed specifications of the preservation content for Earth observation data and derived digital products as an extension to ISO 19165. There has been significant amount of work outside the ISO environment that can be used as a basis for the proposed standard. This work is described briefly below.

3.1. ESA

The European Space Agency (ESA) formed a Long Term Data Preservation (LTDP) Working Group in 2007 for defining and promoting a coordinated approach to preserve and curate European Earth observation space data assets. One of the outputs of this working group was the “Earth Observation Preserved Data Set Content” (EO PDSC), a document providing guidance to data holders on preservation. There have been several versions of this document, the latest having been published in 2012 [9].

3.2. Earth Science Information Partners (ESIP)

Earth Science Information Partners (ESIP) is a U.S. based group of organizations with international membership. It has been active since 1998 and currently has over 180 organizational members including government

agencies, universities and commercial as well as nonprofit entities [10]. In 2011, The Data Stewardship Committee within the ESIP developed an “emerging” Provenance and Context Content Standard (PCCS), which is essentially an enumeration of data and related items that need to be preserved from missions, projects or investigations that support long-term global change research [11]. This list benefitted from the categories of content called for preservation by a report of the US Global Change Research Program’s workshop held in 1998 [12] as well as NASA and NOAA experiences with many Earth observing satellite missions. The purpose of the list is to provide a starting point for developing a standard for content. Thus, the focus is on “what” needs to be preserved, rather than “how”.

3.3. NASA

Based on the PCCS mentioned above, NASA developed its Earth Science Preservation Content Specification (PCS) [13] in late 2011 and has been using it as a requirement for its new missions. For missions that had been in progress or completed before the PCS was developed, it is used as a check list in order to capture and preserve as many of the relevant content items as possible on a best efforts basis. The content items are grouped into the following eight categories: Preflight/Pre-Operations Calibration, Science Data Products, Science Data Product Documentation, Mission Data Calibration, Science Data Product Software, Science Data Product Algorithm Input, Science Data Product Validation, and Science Data Software Tools.

3.4. CEOS WGISS

The Data Stewardship Interest Group within the Working Group on Information Systems and Services (WGISS) of the Committee on Earth Observation Satellites (CEOS) Earth Observation Preserved Data Set Content has adopted the EO PDSC [9] developed by ESA and has evolved it into a more global reference for data preservation [14]. As is the case with references [9], [12] and [13], this document also focuses on “what” needs to be preserved, but also provides guidance on “when” the various content items should be preserved in the course of a mission’s lifecycle.

4. CONCLUSION

Earth observation data and derived digital products are valuable global assets that need to be preserved for the benefit of future generations. To ensure that they remain understandable and reusable decades into the future, it is essential to identify all the associated artifacts that should be preserved along with the data. Frameworks for archival systems, metadata and data quality standards, as well as recent work in national and international agencies provide a good basis for developing a standard that specifies the

content to be preserved. Such a standard will provide uniform guidelines for all organizations around the world involved in Earth observations. Planning for preservation at the beginning of a project or mission will help ensure that the preservation content items are captured at the appropriate times and are not lost as the individuals or groups familiar with those items move on to other activities. This is especially important where several large organizations need to interact for successful implementation of missions, and different teams are responsible for different content items that need to be eventually preserved. Clearly, it is not sufficient that the content items be preserved, but they should also be easily accessible to future users. Establishing persistent identifiers to preserved artifacts and using linked data concepts to connect related items would benefit such access greatly.

5. ACKNOWLEDGMENTS

Ramapriyan's contribution to this paper was supported by NASA's contract with Science Systems and Applications, Inc. Lowe, Mitchell and Murphy contributed to the paper as a part of their duties as employees of NASA.

REFERENCES

- [1] ISO 14721:2012 *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*, <https://www.iso.org/standard/57284.html>.
- [2] ISO 16363:2012 *Space data and information transfer systems - Audit and certification of trustworthy digital repositories*, <https://www.iso.org/standard/56510.html>.
- [3] RLG-NARA Task Force 2007 *Trustworthy Repositories Audit & Certification: Criteria and Checklist*, Version 1.0. Chicago: CRL, February 2007, http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.
- [4] ISO 16919:2014 *Space data and information transfer systems - Requirements for bodies providing audit and certification of candidate trustworthy digital repositories*, <https://www.iso.org/standard/57950.html>.
- [5] ISO 19115-1:2014 *Geographic information -- Metadata -- Part 1: Fundamentals*, <https://www.iso.org/standard/53798.html>.
- [6] ISO 19115-2:2009 *Geographic information -- Metadata -- Part 2: Extensions for imagery and gridded data*, <https://www.iso.org/standard/39229.html>.
- [7] ISO 19157:2013 *Geographic information -- Data quality*, <https://www.iso.org/standard/32575.html>.
- [8] ISO 19165 *Geographic information -- Preservation of digital data and metadata*, <https://www.iso.org/standard/67325.html>.
- [9] ESA:2012 *EO Preserved Data Set Content v4.0*, LTDP-GSEG EOPG-RD-11-0003, July 2012.
- [10] ESIP 2017 <http://esipfed.org/> [Last accessed Sept. 7, 2017].
- [11] ESIP:2011 *ESIP Federation Provenance and Context Content Standard* http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard.
- [12] USGCRP, 1999, Global Change Science Requirements for Long-Term Archiving, Report of the Workshop, October 28-30, 1998, National Center for Atmospheric Research, Boulder, CO. Sponsored by NASA and NOAA, through the USGCRP Program Office, DOI: <http://dx.doi.org/10.7930/J0CZ353N>.
- [13] NASA:2011 *Earth Science Preservation Content Specification* https://earthdata.nasa.gov/files/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf.
- [14] CEOS:2015 *Earth Observation Preserved Data Set Content (PDSC)*, http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf.

SPOT WORLD HERITAGE: SPOT 1-5 DATA CURATION AND VALORIZATION WITH NEW ENHANCED SWH PRODUCTS

Julien Nosavan, Agathe Moreau, Antoine Masse, Benoît Chausserie-Laprée, Claire Caillet

CNES

ABSTRACT

SPOT 1-5 satellites have collected more than 25 million images all over the world during the last 30 years from 1986 to 2015 which represents a unique historical dataset.

Spot World Heritage (SWH) is the CNES initiative to preserve and promote this SPOT archive by providing new enhanced products to users.

A first step has begun in 2015 with the start of the repatriation of the SPOT data hosted in the Direct Receiving Stations spread across the world. From 2017, the SWH initiative is moving into a new operational phase with the launch of the official CNES SWH project and the development of first activities.

SWH processing will take place on CNES High Performance Computing Centre to take advantage of SPOT archive proximity and will use Big Data technologies to manage this volume of data, such as Docker for deployment and Elastic Stack for cataloguing and supervision.

First SWH products are expected to be distributed on CNES Web platforms in 2018 while the whole archive is expected to be processed within 2 years until 2020. Access will be free and controlled as defined in SWH licence agreement.

Index Terms— preservation, SPOT, SWH, long term archive, GeoTIFF, GERALD, reprocessing, curation, processing, Elasticsearch, Docker, Web platform

1. INTRODUCTION

SWH has been announced by France at the GEO Plenary, in January 2014 at Geneva. The announcement included a commitment to make available to the public the archive of the SPOT 1 to 5 satellites images.

Since then, CNES launched a first experiment in partnership with ADS to make SPOT L1C orthorectified products, issued from SPOT 1 to 5 satellite images available freely to the public for non-commercial use: following this work, about 100.000 orthorectified SWH images have been produced at the L1C level by CNES and made available on the French Land products data Centre THEIA (<https://theia.cnes.fr>).

In 2016, SWH has been refocused as a dedicated project and 2 main objectives have been set up: the first objective is to build a new Long Term Archive of SPOT data at level L1A more accessible to users. Indeed, current data are proprietary raw data only accessible through private and commercial

ADS platform and L1A data are GeoTIFF images with radiometric processing applied and geometry of the product kept unchanged.

The second objective of SWH is to provide and diffuse enhanced SWH products in line with ESA Sentinel-2 “standards”, meaning L1B (product with additional corrections and geometric model refined) and L1C (orthorectified product in Top Of Atmosphere reflectance) in order to extend temporal and comparative analysis.

First SWH activities have then started such as the repatriation of remote SPOT data from Direct Receiving Stations to CNES archive system, the extraction of these data on a shared workspace and the integration of the first processing chains on CNES infrastructure.

In 2018, the development of the SWH-PRODCENTER dedicated to SWH activities will follow up and will gather all SWH activities, parallelizing extraction, processing chains and product diffusion to users. First SWH products are expected to be distributed in the end of 2018.

2. SPOT SATELLITES

SPOT 1-5 is a CNES programme, the last satellite of which has now ended its commercial exploitation.

The SPOT satellites main characteristics are the following: each SPOT satellite is composed of two imager instruments (four for SPOT 5) producing squared “scene” images (60 x 60 km), with two modes of acquisition:

- Panchromatic, black and white with a 10m spatial resolution for SPOT 1-4, or 5m/2,5m spatial resolution for SPOT 5 (2,5m with specific THR high resolution mode using 2 images PAN at 5m resolution)

- Multispectral, colour, with a 20m spatial resolution (for SPOT 1-4), or 10m spatial resolution (for SPOT 5) in three bands (green, red and near infrared (SPOT 1-3) or mean infrared (SPOT 4 - 5)).

Table 1 : SPOT main characteristics

Satellite	SPOT 1-3		SPOT 4		SPOT 5				
	HRV		HRVIR		HRG			HRS	
Instrument	PAN	G/R/NIR	PAN	G/R/NIR/MIR	THR	PAN	G/R/NIR	MIR	PAN
Spectral Band									
Resolution	10m	20m	10m	20 m	2,5m	5m	10m	20m	10m

About 25 million SPOT data have been acquired in 30 years, between 1986 and 2015.

3. SPOT CURRENT ARCHIVE

The SPOT satellite images archive is stored in CNES Long Term Archive system STAF (Système d'Archivage et de Transfert de Fichiers) in CNES Toulouse. The STAF is a high capacity archive system based on robotics manipulating magnetic tapes with redundancy and specific operations to ensure the long term archiving.

Despite the 25 million SPOT products referenced, only 7 million are physically stored in STAF because archiving function in SPOT Ground Segment deployed in the Direct Receiving Stations was not automatic. Repatriation is thus actually performed by ADS and CNES to update the STAF archive with about 5 million images still stored in these distant stations. In the end, it is expected to store 12 million images on STAF at the end of the DRS transfer phase, foreseen for the end of 2018. This means that unfortunately, about 13 million data are still stored in Distant Receiving Stations and are not believed – at the moment – to be easily transferable to the STAF (bad storage conditions, corrupted tapes, format compatibility, data losses ...).

The SPOT images are stored as level 0 products under the dedicated GERALD (Generic Exchange for raw Archive Level Data) format, which is constituted as follows:

- One descriptive file (.desc) which describes the acquired segment (or data strip) and its associated scene framing along the Spot Reference Grid (GRS): the .desc file is an ASCII file format

- One or several image files (.ima) which represent the acquired segment: the .ima files are binary files in a proprietary format; they correspond to the first level of processing, completely reversible (auxiliary and ancillary data extractions, quality data analysis and on-board compression for SPOT 5 only).

The characteristics of the GERALD archive format are the following:

- It is the exchange format between the DRS, the former ADS production Centre and the CNES STAF archive,

- It is a long-term format, totally independent from exchange media, self-contained (it contains all the auxiliary and ancillary data to elaborate final products towards users) and without loss on information (it preserves the content of the data as acquired by the satellite and it contains quality indicators to define the original quality of the data).

At last, it shall be noted that the GERALD descriptive file can be formally described using the CNES XIF format thanks to the BEST framework (BEST is a software framework dedicated to data modelling and simulation).

The main drawback of the current GERALD archive format is that it is not directly useable: it is a proprietary format with the image data stored in a raw format, and with on-board characteristics. The level 0 data is not “despatialized”.

4. SWH PRODUCTS

In order to facilitate the use of SPOT products, 3 new SWH products have been defined: L1A, L1B and L1C.

L1A product will be the first image product (GeoTIFF) including basic radiometric corrections and preliminary cloud cover estimation; this product will be principally based on the current SPOT N1A scene format which is the reference for years. This L1A level will replace the current SPOT raw GERALD archive level and will form the new official SPOT archive available on CNES open Web platform with criteria extracted from metadata file (DIMAP).

L1B product will be in segment format and will provide geometric corrections in line with Sentinel-2 L1B product. First of all, inter-bands registration will be reprocessed with optimized L1B ground parameters and geometric model will be refined with Sentinel-2 Global Reference Images and Digital Elevation Model (Planet Observer). On radiometric side, L1B will include new corrections based on first THEIA experiment providing for instance technical masks (water, cloud ...). L1B will also include new algorithms optimized and ready to deal with huge volume of data; in particular, a new denoising algorithm based on Non Local Bayes technique [1] has been developed and optimized for SPOT 5 THR mode production with reduction of computational complexity and a tile-ready processing. Example of this SPOT THR product is shown in Fig.2:

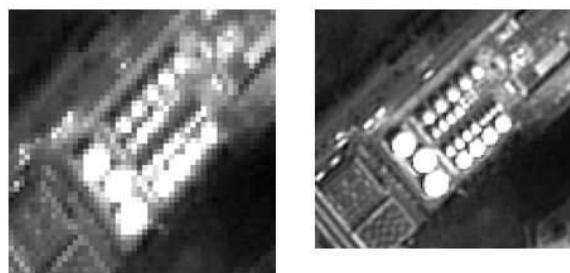


Figure 2: Example of SPOT THR mode result: (left) SPOT 5 HMA image (5m) and (right) SPOT THR product (2.5m)

L1C product will be the orthorectified product in Top Of Atmosphere reflectance, still in line with Sentinel-2 L1C product with a specific split in 20km x 20km tiles (sub-tiling of 100 km x 100 km S2 L1C tiles).

All these products will be based on the current SPOT format (including DIMAP metadata file) and will also include evolutions to take into account Sentinel-2 specificities.

5. SWH SOFTWARE ARCHITECTURE

SWH software architecture relies on a strong reuse of existing tools to minimize development costs and secure validation phases regarding to the volume of data to process. SWH development is organized around a main Centre dedicated to SWH activities: the SWH-PRODCENTER.

This Centre gather all SWH activities delegated to SWH components managed with the help of a centralized database SWH-DBREF based on Elasticsearch to deal with the volume of the data to index. Supervision will be ensured by

Kibana software that natively allows the setting of customizable “views” and “dashboards” for operators. It is described in Fig.3:

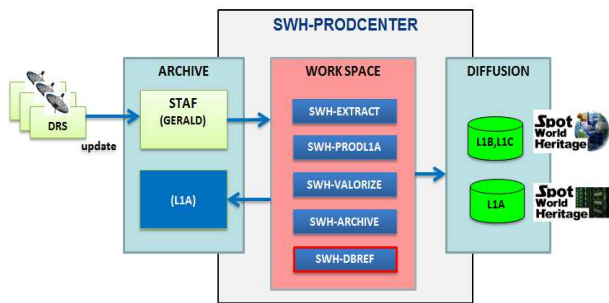


Figure 3 : SWH-PRODCENTER

SWH-EXTRACT is the data extraction chain based on an internal CNES development using itself CNES internal tools (VDLIB, STAF client). It is described in Fig.4:

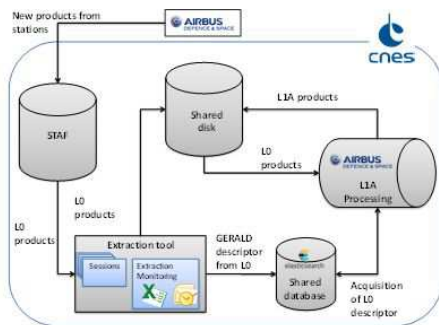


Figure 4 : Extraction and processing flow

The objective of the tool is to extract GERALD data from the STAF to a shared SWH workspace. The tool notifies the LIA processing chain of the availability of the GERALD by adding indexes in SWH-DBREF. The tool can be configured to extract all the SPOT data archive and can be easily monitored through the extractions reports and e-mails. SWH-PRODLIA is the LIA processing chain based on the integration and the automation of the operational ADS SPOT N1A processing chain on CNES infrastructure to take advantage of SPOT archive proximity and CNES High Processing Centre. It is described in Fig.5:

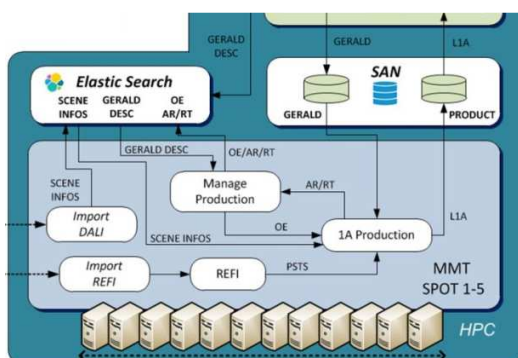


Figure 5: SWH-PRODLIA processing chain

The objective of the tool is to generate LIA products from GERALD products using the shared SWH workspace. The tool notifies the availability of the product by adding indexes in SWH-DBREF. LIA processing chain is scalable and will use Docker software to deploy Docker images on CNES HPC servers to be – when necessary - independent of the system environment.

SWH-VALORIZE is the LIB/LIC processing chain based on the reuse of CNES MUSCATE software already used in THEIA experiment. It is described in Fig.6:

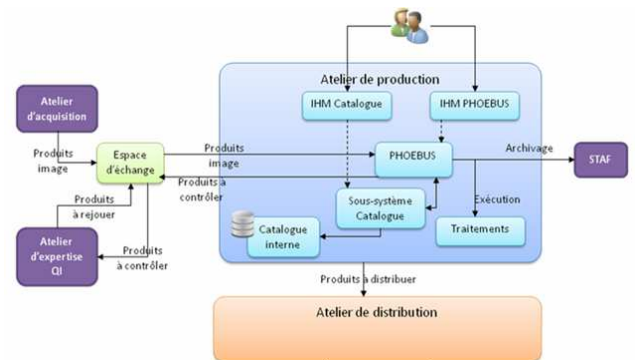


Figure 6: MUSCATE functionalities

The objective of the tool is to generate and diffuse LIB and LIC products from LIA products using the shared SWH workspace. The tool notifies the availability of the products by adding indexes in SWH-DBREF. First THEIA SWH processing chains based on CNES SIGMA software for geometric refining and orthorectification will also be reused and adapted to SWH/Sentinel-2 specificities. Finally, LIB and LIC products will be diffused on a dedicated archive catalogue. SWH-VALORIZE will extract some of these metadata to index the LIB and LIC products and provide search criteria to the users. A cartographic search will also be provided like THEIA platform.

SWH-ARCHIVE is the archiving processing chain based on the reuse of CNES REGARDS software. It is described in Fig.7:

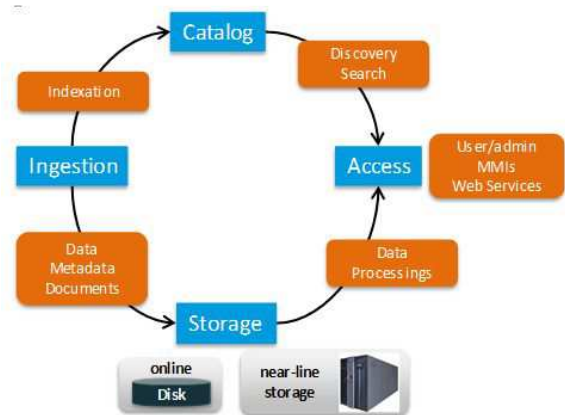


Figure 7: REGARDS functionalities

REGARDS is a new CNES product currently in development. It is based on micro-services architecture with external interfaces using HTTP Restful services. REGARDS is a generic product which means it will be used by several missions in several data centres (earth observation, astronomy, space sciences, etc.). For each project a new implementation will be deployed allowing project specificities. For SWH, REGARDS will be configured and a specific plugin will be developed to interface SWH-DBREF to identify which L1A products are to archive. The L1A product will be a zip containing the GeoTIFF images as well as a DIMAP descriptive file and additional information (cloud coverage, traceability ...). REGARDS will extract some of these metadata to index the L1A product and provide search criteria to the users. A cartographic search will also be provided by REGARDS. Once the L1A product is archived, the associated GERALD will be deleted from the online storage.

The objective of SWH-ARCHIVE is to archive back L1A products in replacement of GERALD archive from the shared SWH workspace to STAF. The tool notifies the archiving status by adding indexes in SWH-DBREF.

6. SWH CHALLENGES AND IT ARCHITECTURE

The current estimation of 12 million images SPOT represents a volume of ~700 TBytes for GERALD data, 1 PByte for L1A products, 1 PByte for both L1B and L1C products (using compression).

GERALD extraction is expected to be performed during one and a half years depending on the timing of SPOT data retrieval from the Direct Reception Stations.

L1A processing capacity has been estimated with benchmarks on ADS SPOT N1A processing chain and a cluster of 17 nodes (24 core, 128 Go RAM) has been identified to generate the whole L1A archive in one year meaning a daily production of 90.000 scenes.

L1B/L1C processing capacity has been estimated with current performances observed on THEIA prototype and ongoing optimizations on CNES MUSCATE software. 60% of L1A expected to be processed in L1B and L1C level, depending of the Cloud coverage and the Quality image configuration put in place in THEIA processing. The whole L1B and L1C products are then expected to be processed in less than 2 years.

L1A archiving is expected to be performed during one and a half years but

All these components will be launched in background with a weekly configuration (TBC) and a daily supervision. However, priority can then be taken into account if a specific area is asked for, depending on the partnership put in place.

SWH IT architecture is based on the use of ~20 nodes (24-core processors) and optimized GFPS shared workspace for computation. A final diffusion workspace will host data for the SWH Web platform.

Here is displayed in Fig. 8 the global IT architecture of SWH-PRODCENTER built on CNES facilities:

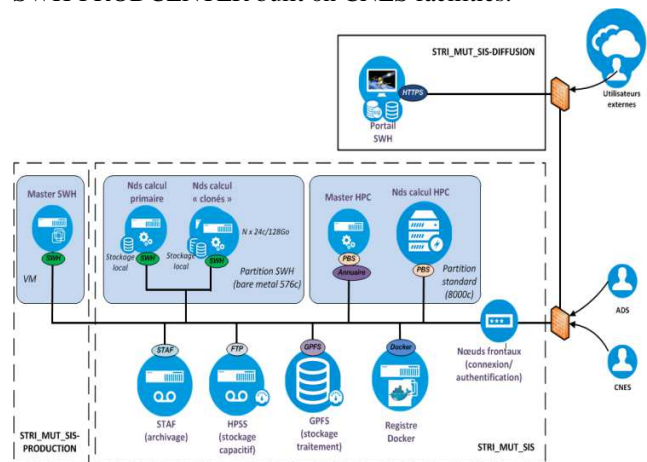


Figure 8: SWH IT architecture

7. SWH SCHEDULE AND PRODUCT DIFFUSION

SWH-EXTRACT software is validated and operations have started in August 2017: GERALD data are currently being extracted.

L1A processing chain development has started in July 2017 and the operational version is expected in summer 2018.

All remaining activities (SWH-VALORIZE, SWH-ARCHIVE and SWH-DBREF) are gathered in the SWH-PRODCENTER development that will start in the beginning of 2018.

Production is expected to start in the end of 2018 and first products are then also expected to be distributed in the end of 2018 while the whole archive is expected to be processed within 2 years until 2020.

SWH products will be accessible through two dedicated Web platforms: one dedicated to L1A archive products, one dedicated to L1B and L1C products.

SWH licensing is being defined but the main goal is to provide a free and controlled access to all SPOT data. The new generated SWH products will then be accessible free of charge to registered users.

8. REFERENCES

- [1] A. Masse, S. Lefèvre, R. Binet, S. Artigues, P. Lassalle, G. Blanchet, and S. Baillarin, "Fast and accurate denoising method applied to very high resolution optical remote sensing images", Proc. SPIE 10427, Image and Signal Processing for Remote Sensing XXIII, 1042703 (4 October 2017), doi:10.1117/12.2277705.

20-YEARS OF ESA SPACE SCIENCE DATA ARCHIVES MANAGEMENT

*Christophe Arviset¹, Deborah Baines², Isa Barbarisi³, Sebastien Besse⁴, Guido de Marchi⁵,
Beatriz Martinez⁶, Arnaud Masson⁷, Bruno Merin¹, Jesús Salgado², Claire Vallat⁶*

¹ESAC Science Data Centre, ESA, Madrid, Spain

²ESAC Science Data Centre, QUASAR for ESA, Madrid, Spain

³ESAC Science Data Centre, SERCO for ESA, Madrid, Spain

⁴ESAC Science Data Centre, AURORA for ESA, Madrid, Spain

⁵ESAC Science Data Centre, ESA, Noordwijk, The Netherlands

⁶ESAC Science Data Centre, RHEA for ESA, Madrid, Spain

⁷ESAC Science Data Centre, TPZ VEGA for ESA, Madrid, Spain

ABSTRACT

In the mid-90s, ESA decided to change its data management strategy and started to build at ESAC (European Space Astronomy Centre), data archives for its space science missions, initially for its Infrared Space Observatory and then expanding through other astronomy missions and later on, to planetary and solar helio physics missions. The ESAC Science Data Centre now hosts more than 15 science archives, with various others in preparation.

Technology has evolved a lot through this period, from the simple web pages towards rich thin layer web applications, inter-operable and VO built-in archives. Maintaining old legacy archives while building new and state of the art ones (eg Gaia), managing people and preserving expertise over many years, offering innovative multi missions services and tools to enable new science (ESASky) have been some of the many challenges that had to be dealt with.

Future prospects ahead of us also look exciting with the advent of the "Archives 2.0" concept, where scientists will be able to work "within" the archive itself, bringing their own software to the data, sharing their data, code and results with others.

Data Archives have been and continue to be in constant transformation and they are now evolving towards open and collaborative science exploitation platforms..

Index Terms— Archives, Virtual Observatory, Space missions, Long Term Data Preservation

1. ESAC SPACE SCIENCE ARCHIVES: AN EVER GROWING FAMILY

In the past, ESA was leaving to the scientific community the role of archiving its data holdings (for example for IUE or EXOSAT), In the mid-1990s and with the advent of the WWW offering new possibilities for data searches and dissemination, ESA changed its strategy and decided to host on-line archives with for its space science missions. This started with ISO Data Archive in 1998.

Based on its success, it was decided to re-use the existing expertise and to develop the XMM-Newton Science Archive (released in 2002). Herschel and Planck archives were ready by launch (2009), representing an important part of the missions' Science Ground Segment. EXOSAT and SOHO archives were also added in 2009, using new Java development standards.

Together with the expansion of ESAC activities towards ESA planetary missions, the Planetary Science Archive (PSA) was built, hosting all ESA planetary data (initially Giotto (2004), Mars Express (2005) and Huygens (2006)). As all these missions were using the same data format (PDS for Planetary Data System), it was decided from the start to consolidate them all into a multi mission archive. The PSA was later on enriched with data from Venus Express (2009), SMART-1 (2010), Rosetta (2010) and now Exomars TGO (2016).

Consolidation of ESA space science archives to ESAC continued with the migration to ESAC of the European HST Archive from ESO (2012) and the Ulysses and Cluster archives from ESTEC (2013).

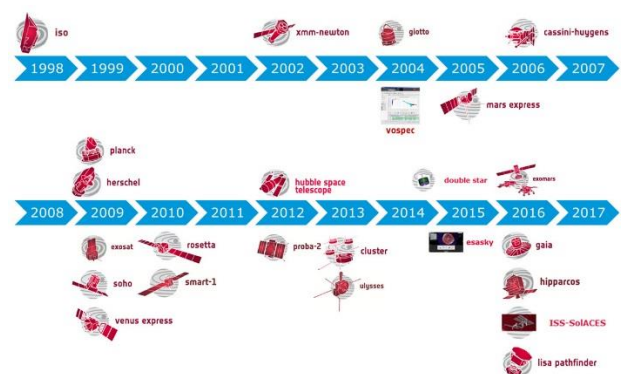


Figure 1: ESAC Space Science Archives

Building on the wide variety of astronomy archives at ESAC, a major milestone was reached in 2015 with the release of ESASky, bringing all ESA astronomical data

holdings (plus some others) through an innovative and user friendly sky viewer.

In 2016, Gaia Data Release 1 was made public, including most of the Hipparcos catalogues, as well as the archive from LisaPF.

Overall today, ESAC Science Data Centre (<http://archives.esac.esa.int/>) hosts a dozen of archives containing science data from over 20 space science missions.

2. ESAC SCIENCE ARCHIVES STRATEGY

In 2012, taking into account the vast and wide variety of ESA scientific data holdings available, we defined the ESAC Science Archives long term strategy articulated around three main pillars:

1. Enable Maximum Science Exploitation
2. Enable efficient long-term preservation
3. Enable cost-effective archive production by integration in projects

The science data is the ultimate delivery of any ESA space science mission. One of the metrics used to determine the success of a mission is the number of refereed scientific publications in the literature. Therefore, the archive must provide the best science data together with all the necessary services and tools to maximize its science exploitation. ESA has the responsibility to ensure that the data hosted in its archive are of the highest quality, scientifically validated or even peer-reviewed in some cases, hence fully reliable and with the associated level of documentation, to allow scientists to do their science and then write their scientific papers. The release of ESASky in 2015 was a major milestone in this respect, enabling science exploitation from multi missions' data [1].

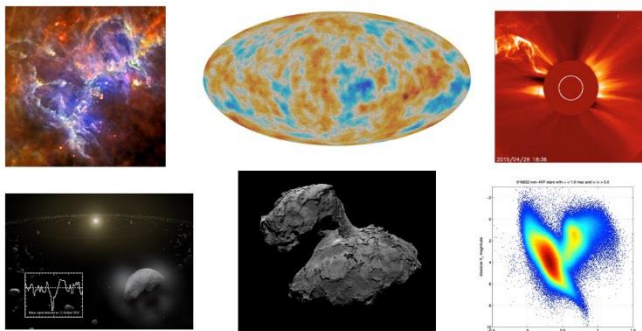


Figure 2: Data from ESA Space Science Missions

The archives must remain available for a long time, much longer than the current mission lifetime which already typically spans over 15 years. ESA has a commitment to preserve in the long term not only the data and associated services to access these, but as well the knowledge about this data. By consolidating all ESA space science archives in one place under the umbrella of the ESAC Science Data Centre, we can ensure strong re-use of technology and people expertise across archive projects. People working on

active archives can also maintain the legacy archives, which also brings cost savings. Additionally, recognizing that the IT technology evolves rapidly while archive systems must perdure for many years, technology migration for archives will be required every five to seven years to ensure state of the art services.

ESA Science Operations have traditionally been organized by individual missions whereas the archive development, operations and maintenance is a transversal service to all missions. In the past, archives were often developed only in the final stage of the operations of a mission. Nowadays, with even more distributed and complex Science Ground Segment systems, the archive becomes sometimes the heart of the overall system (eg for Euclid) and therefore needs to be developed in the very early phases of the missions.

3. ARCHIVES DESIGN TECHNOLOGY EVOLUTION

Architectural design does matter a great deal when building an archive system that must be preserved for decades. Modularity and flexibility are key concepts in archive systems architecture, to facilitate technology evolution through time. The right balance needs to be found between providing state of the art services with newer technology options (including migration of existing systems) and avoiding the technology buzzes that won't survive long.

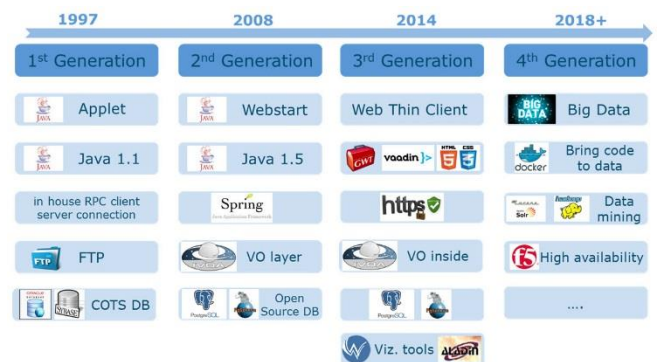


Figure 3: ESAC Archives Technology Evolution

Over the last 20 years, ESAC archives have gone through important migrations. RDBMS systems went from COTS (Sybase, Oracle) towards open source solutions (PostgreSQL), enriched by discipline specific excellent plugins (eg pgSphere and PostGIS), and new databases systems are now being investigated (PostgresXL, CitusData, see reference [2]) to address big data challenges brought by new missions (Gaia [3], Euclid).

An application server allows to separate the data and metadata from its presentation and also others many other functionalities (caching, security, activity login, etc...). While we had to develop our own software in the early days, we then used existing frameworks, such as Spring and

Hibernate which are IT industry standards and therefore facilitate greatly the development and maintenance.

On the GUI side, while Java was the best choice in the late 90s to build rich GUI interfaces, its support became poorer and poorer, while other web technologies become more advanced. In the late 2000s, we decided to migrate all our archives GUIs towards web thin clients, providing faster loading, no Java installation and overall better browser support. We chose GWT to continue benefitting from our experience Java software developers.



Figure 4: Evolution of ESAC Archives GUIs

To ensure interoperability with other archives, we have been developing a VO (Virtual Observatory) layer on top of the existing archive APIs, building VO services through standard protocols (in particular VOTable, Simple Image Access Protocol, Simple Spectra Access Protocol, Simple Line Access Protocol) and connecting to external VO Tools through SAMP (Simple Application Messaging Protocol). With more advanced archives such as Gaia and Euclid, we started to directly use the VO protocols (eg Table Access Protocol, Universal Worker Service, VOSpace) to other synchronous and asynchronous archive services. In this context, Gaia is definitely the first VO-built-in archive.

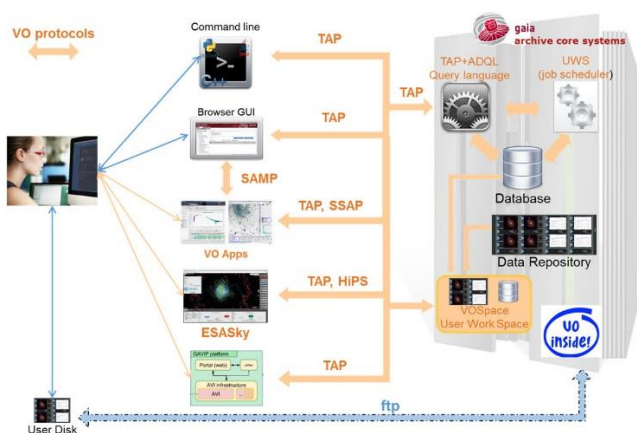


Figure 5: Gaia Archive Architecture

It is also interesting to note that some of the VO protocols (eg TAP, SAMP), initially designed for astronomical data are being used for archives in other scientific disciplines (planetary and solar heliophysics).

4. TOWARDS ARCHIVES 2.0 PARADIGM

The traditional way scientists usually interact with the archives can be called the "bring the data to the user" concept. Scientists go to the on-line archive, perform queries to determine which data they want, usually supported by some light weight visualization tools, and then download the data to their computer. From there, they use standard data analysis packages or their own scripts to analyze the data further and then later on write their scientific papers.

New missions are bringing unprecedented amount of data, in the order of hundreds of GBytes or even PBytes. This calls for new models to access and interact with the data.

First, querying billions of data holdings and cross matching them with other catalogues might require longer than what is expected for an interactive query session, hence the need to provide asynchronous services where complex queries can be queued, executed in the background and then provide results to the scientist after a few minutes. Results of such queries can still contain hundreds of millions of results and might be better stored (and indexed for performance) into the archive so the scientist can use it for further refinement.

Second, the scientist cannot download anymore all the data to her computer, as this would take too long and she probably would not have enough disk space anyway. It is up to the archive to provide user workspaces both for database (for user tables as seen above) and for data storage (done through VOSpace for example), so the data does not need to be transferred over the network.

When the data reside in the user workspace in the archive itself, the scientist wants to run standard data analysis package or her own software and scripts onto her data. This is the new archive concept "bring the code to the data". Most probably, archive data centres will also have to provide computing facilities next to their data so archive users can work with the data where the data actually resides. This could be done through dedicated cloud hardware infrastructure at the data centre itself (or eventually an hybrid solution involving external clouds if some data can also be copied there). New technologies (eg Docker containers, Jupyter notebook) should facilitate this implementation and initial examples look very promising.

This new concept of "Archive 2.0" would also allow scientists to collaborate much more easily. Users could share their workspaces (database table, data storage, but as well their own software and scripts) with other archive users.

We think that from the original metadata and data repositories, the archives are evolving towards open and collaborative science exploitation platforms.

5. ESASKY: BIG DATA VIZUALIZATION

As part of ESA strategy to increase science exploitation of its data holdings, the ESAC Science Data Centre built a completely new tool, called ESASky (<http://sky.esa.int/>). ESASky is a new science-driven discovery portal for most ESA astronomical missions that gives users worldwide a simplified access to high level science ready products from ESA and other data providers. The tool features a sky exploration interface and a single/multiple target search interface. It does not require any prior knowledge of the specific details of each mission. Users can explore the sky in multiple wavelengths, quickly see the data available for their targets and retrieve transparently the relevant science products from the corresponding archives.

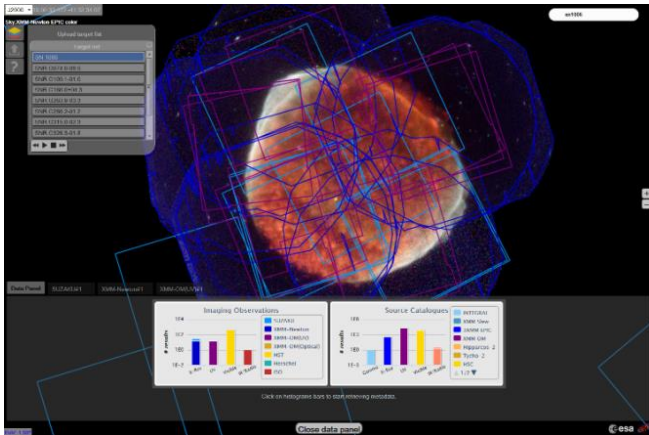


Figure 6: ESASky discovery portal

ESASky is making full use of protocols that have been developed within the IVOA (International Virtual Observatory Alliance), which enable interoperability between astronomical data. On the client side, visualization is made fast by using Hierarchical Progressive Survey (HiPS), which splits the sky into various levels of tiles (depending of the mission) to minimize data transfer from the server to the client. Mission coverage (footprints) is described with MOC (Multi-Order Coverage). On the server side, other techniques like TAP services on common data models for fast and performant searches, database geometrical indexes, internal connections between databases and wrappers around the individual mission archives to download the final science data have been setup to allow the handling of big amounts of data in a simplified way.

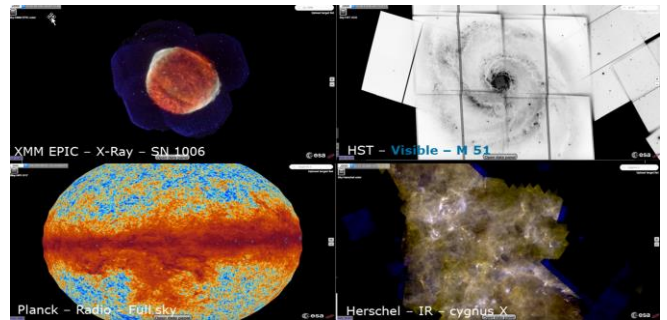


Figure 7: ESASky, see the sky with different "eyes"

6. CONCLUSIONS

Since the first public release of the ISO Data Archive in 1998 to the most recent Gaia archive release in September 2016, the ESAC Science Data Centre has converted itself into ESA's digital library of the universe, presenting and preserving reliable space science data for over twenty scientific missions. ESA Space Science Archives strategy is clearly articulated towards maximizing the science exploitation of data, ensuring long term preservation of data, knowledge and software, while supporting the development and operations of the Science Ground Segments.

This can be achieved through very close integration of scientists and software engineers, ensuring archives are science driven, and supported by strong IT expertise that need regular technology migration through time.

To cope with new archive challenges (open data, big data volume, need to bring the code to the data, open and collaborative archives), a new paradigm for archive development and archive users is ahead of us that will bring the archives towards an exciting era that will revolutionize the way scientists interact with data..

7. REFERENCES

- [1] B. López Martí, B. Merín, F. Giordano, D. Baines, E. Racero, J. Salgado et al., "ESASky: The whole of space Astronomy at your fingertips", *Proceedings of the XII Scientific Meeting of the Spanish Astronomical Society*, July 2016, [arXiv:1610.09826](https://arxiv.org/abs/1610.09826).
- [2] P. de Teodoro, S. Nieto, J. Salgado, C. Arviset, "Considering scale out alternatives for big data volume databases with PostgreSQL" *BIDS 2017 conference*.
- [3] A. Mora, J. González-Nuñez, D. Baines, J. Durán, R. Gutiérrez-Sánchez, E. Racero, J. Salgado, JC Segovia, "The Gaia Archive" *Proceedings IAU Symposium No. 330*, 2017, [arXiv:1706.09954](https://arxiv.org/abs/1706.09954).

CLOUD BASED EARTH OBSERVATION DATA EXPLOITATION PLATFORMS

A. Romeo(1), S. Pinto (1), A. Marin (2), Sveinung Loekken (3)

(1) RHEA Group, (2) Solenix, (3) European Space Agency

ABSTRACT

In the last few years data produced daily by several private and public Earth Observation (EO) satellites reached the order of tens of Petabytes, representing for scientists and commercial application developers both a big opportunity for their exploitation and a challenge for their management. New IT technologies, such as Big Data and cloud computing, enable the creation of web-accessible data exploitation platforms, which offer to scientists and application developers the means to access and use EO data in a quick and cost effective way.

RHEA Group is particularly active in this sector, supporting the European Space Agency (ESA) in the Exploitation Platforms (EP) initiative, fostering development of technology to build multi cloud platforms for the processing and analysis of Earth Observation data, and collaborating with larger European initiatives such as the European Plate Observing System (EPOS) and the European Open Science Cloud (EOSC).

Of these technologies, Thematic Exploitation Platforms build virtual work-spaces for scientists to access and process EO data, collaborate and share results, while solutions like the multi cloud EO data processing platform aims to integrate ICT resources and EO data from different vendors in a single platform.

This work will present an overview of the TEPs and the multi-cloud EO data processing platform, and discuss their main achievements and their impacts in the context of the distributed research infrastructures such as EPOS and EOSC.

Index Terms— Earth Observation, Platform, Data Analysis, Data Exploitation

1. INTRODUCTION

Currently Earth observation private and public satellites produce tens of Terabytes of data per day and the trend is increasing. If such huge amount of data represents for scientists and commercial application developers' big opportunity for their exploitation, on the other hand is a challenge for their management. Latest IT technologies, such as Big Data and cloud computing, via the creation of web-accessible data exploitation platforms, allows scientists and application developers to access and use EO data online in a quick and cost effective way, without the need to download the data locally.

RHEA Group is currently supporting several of such data exploitation platforms activities, such as the European Space Agency (ESA) Exploitation Platforms (EP) initiative. RHEA is also supporting technology development to build multi cloud platforms for the processing and analysis of Earth

Observation data, and collaborates, together with ESA, to large European initiatives in the framework of the European Strategy Forum on Research Infrastructures (ESFRI), such as the European Plate Observing System (EPOS), and the European Open Science Cloud (EOSC).

2. EXPLOITATION PLATFORM CONCEPT

An EP is a virtual workspace, providing a user community with access to

- large volume of data, both EO and not EO as relevant for the specific use
- on platform and on user premises algorithm development and integration environment
- interactive and batch processing software and services (e.g. toolboxes, visualization routines)
- computing resources
- collaboration tools (e.g. forums, wiki, etc.)



FIGURE 1 EXPLOITATION PLATFORM CONCEPT

All these features are provided in a cloud based environment which has the benefits to:

- removing the ICT ownership costs from data users. They can access data and ICT resources directly on the platform without the need to download data and run their algorithm on their own hardware. It also allows elasticity in resource allocation which would be otherwise impossible to achieve.
- improve the usage of EO data in areas where internet connectivity is weak. In fact, as EO data can be processed and analysed on the platform which is accessible via a web browser, the amount of data transferred from the platform to the user is drastically reduced if compared with the classical scenario in which data need first to be downloaded.

- simplify information sharing as all processing and analysis result are already online. Scientists can easily provide links to their results or create shared work spaces with their colleagues

3. ESA THEMATIC EXPLOITATION PLATFORMS

When an EP is dedicated to a specific Theme, it becomes a Thematic Exploitation Platform (TEP). Currently, ESA has seven TEPs [1] dedicated to:

- geo-hazards monitoring and prevention (GEP)
- coastal zones (C-TEP)
- forestry areas (F-TEP)
- hydrology (H-TEP)
- polar regions (P-TEP)
- urban areas (U-TEP)
- food security (FS-TEP)

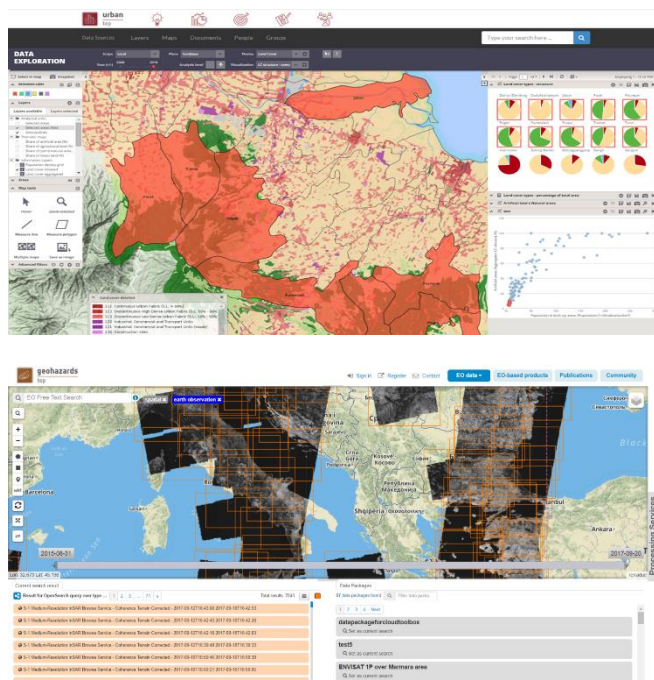


Figure 2 TEP INTERFACE EXAMPLES. TOP: U-TEP. BOTTOM: GEP

The first six TEPs have been developed in the course of the last 2 years and now are in their pre-operational phase and have already several pilots and success stories implemented, such as, for example:

- Support to fish farming, where Coastal TEP is powering the Supporting our Aquaculture and Fisheries Industries (SAFI) services, analyzing several water quality and temperature variables generated from satellite data to pinpoint suitable aqua-farming locations around the globe [2]
- Human footprint products, such as the Global Urban Footprint, with yearly evolution and city greenness

analysis products, generated by Urban TEP and used by World Bank and other pilot users [3]

- Support to operational ice-patrol and iceberg detection services via satellite data analysis performed on the Polar TEP [4]
- Systematic analysis of seismic zones, with the generation of interferograms and coherence maps on the GeoHazards TEP, monitoring an area of three million square kilometres in 200 m blocks [5]

They have been also involved in various degrees in international initiatives. For instance, both GEP and C-TEP provide a prototype for the International Disaster Charter. GEP is involved also in the CEOS WGD disaster Volcano and Earthquake pilot, EPOS Satellite Data Thematic Core Services and via the latter in the European Open Space Cloud.

The Food Security TEP project started in April 2017 and its development is based on Agile and dev-ops approach so it is going to deploy in pre-ops its first release in November but already has some collaboration in place with WFP, FAO and commercial entities in the sector.

Each TEP tackle its thematic community which internally may be extremely varied. In fact it may span from scientists to commercial operator, from EO data expert to decision makers. TEPs different GUI and the plethora of tools available are a consequence of providing to each user category the information and instruments to perform their activity in the most effective way.

Even if the theme tackled by a TEP is well defined, in many cases interactions are requested between the different TEPs in order to achieve a specific goal. For instance, FS-TEP may need information on coastal zones and inland wetland which can be provided by C-TEP and H-TEP in order to efficiently provide services on specific aquaculture sites. Therefore, interoperability between TEPs and also other platforms is the next step of the evolution of the Exploitation Platform program which aims to create a network of data and service providers to further streamline the usage of EO data

4. MULTI-CLOUD EO PROCESSING PLATFORM

In parallel to the engineering and management support to ESA for the development of TEPs, RHEA Group developed its own solution to enable the processing of Earth Observation data on multi cloud platforms. The Multi-cloud EO processing platform provides the technology to integrate ICT resources and EO data from different vendors in a single platform. It then become the building block for creating data exploitation platforms which transparently incorporate the offerings of different underlying cloud platforms and has the advantage of avoiding cloud vendor lock-in.

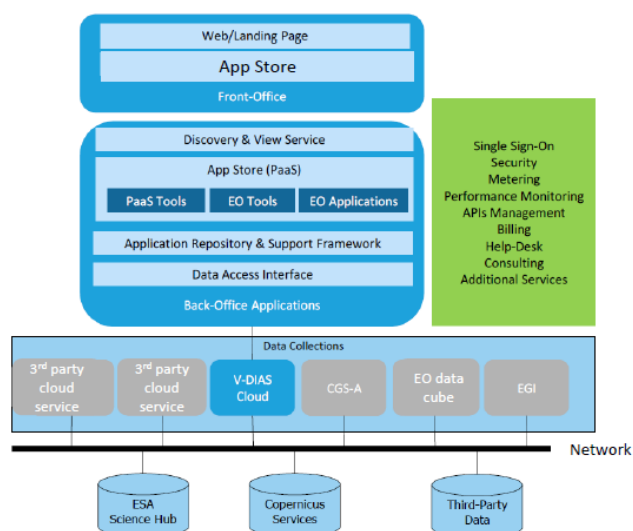


FIGURE 3 MULTI CLOUD PROCESSING PLATFORM LOGICAL ARCHITECTURE

The Multi-Cloud EO Processing Platform offers 3 main services:

- **Multi-cloud data discovery:** a data discovery service allows to discover and select EO scenes relevant to an EO application. It uses a metadata catalogue currently tracking nine satellites including Landsat-8, Sentinel-1 and 2, and currently able to hold up to 10+ millions of scenes.
- **Multi-cloud data management and access:** a data management service provides a data location API that allows to expose and stage EO data on the cloud infrastructure on which the EO application is deployed and executed. It uses a global high-performance data management system providing access to distributed storage resources.
- **Multi-cloud application deployment:** an application deployment service allows to select the cloud infrastructure on which to deploy the EO application, and provides automated deployment, execution and monitoring. It uses a multi-cloud application management service automating the full application management lifecycle, and providing connectors to multiple clouds, public and private, leveraging both open source and proprietary cloud APIs.

The platform has been implemented by the integration of 4 main products:

- SlipStream: a multi-cloud broker by SixSq, a RHEA Group company, which allows to define cloud independent application deployment recipes which can then be used transparently in different cloud platforms
- ONEDATA: a global data management system by Cyfronet. It allows to create a virtual storage

volume which integrate data in different locations, and different cloud environment which are presented as they were in a single local volume

- SatCat: a visual EO product catalogue by EOproc
- Nuvla: a web front end for SlipStream by SixSq, which offers also an App Store for the discovery of the application available in the platform

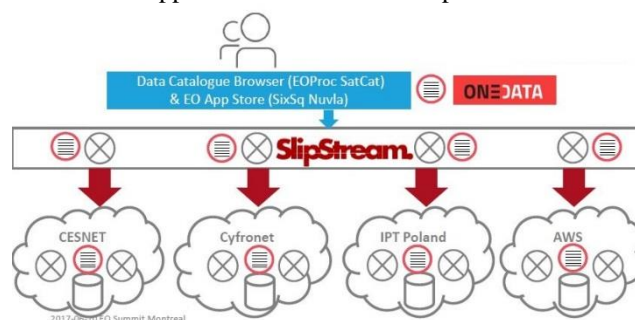


FIGURE 4 MULTI CLOUD PROCESSING PLATFORM MAIN COMPONENT

The Multi-cloud EO processing platform has been demonstrated with the EGI Federated Cloud, Innovation Platform Testbed Poland and the Amazon Web Services cloud.

The concept is currently evolving with the integration of the Open Data Cube in the data management and access service. This will provide the platform the capability to handle more easily data analytics tasks using the feature of the data cube and still maintaining a high level of abstraction for the users.

REFERENCES

- [1] <https://tep.eo.esa.int/>
- [2] http://www.esa.int/Our_Activities/Observing_the_Earth/Fish_farms_guided_by_Sentinels_and_the_cloud
- [3] http://www.esa.int/Our_Activities/Observing_the_Earth/New_map_offers_precise_snapshot_of_human_life_on_Earth
- [4] http://www.esa.int/Our_Activities/Observing_the_Earth/Iceberg_patrol_gains_faster_updates_from_orbit
- [5] http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-1/Keep_an_automatic_eye_on_seismic_zones

VIRTUAL EXPLOITATION ENVIRONMENT DEMONSTRATION FOR ATMOSPHERIC MISSIONS

Stefano Natali⁽¹⁾, Simone Mantovani⁽¹⁾, Gerhard Triebnig⁽²⁾, Daniel Santillan⁽²⁾, Marcus Hirtl⁽³⁾, Barbara Scherllin-Pirscher⁽³⁾, Cristiano Lopes⁽⁴⁾

⁽¹⁾ SISTEMA GmbH, Vienna, Austria

⁽²⁾ EOX IT Services, Vienna, Austria

⁽³⁾ Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Vienna, Austria

⁽⁴⁾ ESA ESRIN, Frascati, Italy

ABSTRACT

The scientific and industrial communities are being confronted with a strong increase of Earth Observation (EO) satellite missions and related data. This is in particular the case for the atmospheric sciences communities, with the Copernicus Sentinel-5 Precursor satellite as well as upcoming Sentinel-5, -4, and ESA's Earth Explorers scientific satellites ADM-Aeolus and EarthCARE. The challenge is not only to manage the large volume of data generated by each mission / sensor, but also to allow users to analyse the data streams in near-real-time and for long-term monitoring tasks. Creating synergies among the different dataset will be key to exploit the full potential of the available information.

As a preparation activity supporting scientific data exploitation for Earth Explorer and Sentinel atmospheric missions, ESA funded the "Technology and Atmospheric Mission Platform" (TAMP) [1] [2] project, with the twofold aim of demonstrating (1) that multiple data sources (satellite-based data, numerical model data, and ground measurements) can be simultaneously exploited by users (mainly scientists), and (2) that a fully Virtual Research Environment (VRE) that allows avoiding the download of all data locally, and retrieving only the processing results is the optimal solution.

With the "Virtual Exploitation Environment Demonstration for Atmospheric Missions" (VEEDAM) project, the concept of VRE is further extended: data visualization capabilities have been improved providing volumetric data slicing tools; moreover a Jupyter notebook interface has been deployed, providing the users with effective data access service and flexible processing tools. This paper aims at outlining the implemented concept for heterogeneous geospatial data, highlighting the efforts devoted to facilitate the user experience in both data visualization and data exploitation (processing, cross-correlation, validation). Use cases show how specific events (e.g. volcanic eruptions) can be easily studied within the VRE simultaneously exploiting all available information.

Index Terms— VRE, Atmospheric Sciences, Sentinels, data exploitation

1. THE CONCEPT

The ambitious scope of the VRE is to support the scientific and technical communities to access and use past, current, and future atmospheric science data. To this aim, TAMP has been developed to simultaneously access, visualize, correlate, and download EO-based products, numerical model data, and reference/validation datasets (data triangle).

In order to achieve the final goal, the platform implements a fairly generic geospatial data management technology, together with a powerful data visualization engine and an effective and easy-to-use data processing interface. The full platform is deployed as a VRE installed at the Austrian Meteorological Service (ZAMG) premises, fully virtualized and accessible by the user via web browsers or via secure terminal connections.

The following services are provided to the users:

- Social-like portal
- Data access and view services
- Processing / data assessment services
- Download (data, plots) service

With TAMP, the user interacts in real time with long time series of a large variety of data, visualizes and processes the data without the need of downloading a single bit in the local environment. Once done, the user can decide whether or not to retrieve only the final results.

2. ARCHITECTURE

The TAMP platform features two main layers: the user-interface layer and the data management layer (see Fig. 1). Within the data management layer, the Data ARchive module (DAR) is the only I/O interface toward the "data storage area", enabling WCS2.x services thus allowing multi-temporal multi-dimensional multi-field queries. Two modules are further deployed on the data management layer: the processing resources and the Jupyter virtual machine. The processing resources are as close as possible to the data, and are triggered by the Data Analysis and Visualization Environment (DAVE), by the user shell and the Jupyter notebook, optimising the resources usage and libraries sharing (the user has the same processing capabilities from each of the interfaces).

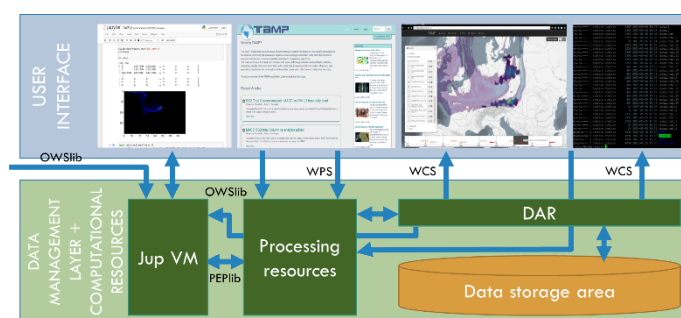


Fig. 1. TAMP overall architecture

The user interface layer exposes four access modes:

- The Portal Information Page (PIP) is the platform landing page (<http://vtpip.zamg.ac.at/>): it implements a user-centric approach where each user can upload owned collections, access the DAVE and the Jupyter interfaces, publish and share results with other users and social networks.
- DAVE is the graphic interface to perform multi-collections multi-temporal data visualization, to apply pre-defined processing utilities, and to download results in terms of data and plots.
- The Jupyter notebook works as a Python console and, besides making available the TAMP processing functions, allows implementing and executing new processors and downloading the results.
- The Command Line Interface (CLI) is a real Linux virtual machine, with direct access to the data and processing functions (deployed as a Python library), where user algorithms can be installed, executed, and the resulting collections can be re-ingested into the platform to be visualized and analysed via the other (DAVE, Jupyter) modules.

3. USER EXPERIENCE AND USE CASES

TAMP has been designed to be user-centric, so each user with credentials can access the whole set of user interfaces. In order to facilitate the user experience, a wiki page has been created (<http://vtpip.zamg.ac.at/wiki/>) and a set of short videos (around 1 minute each) shows how to perform basic and advanced operations.

In the following sub-sections two uses cases are presented to demonstrate the flexibility of the platform to enable long time series analysis of heterogeneous data (satellite and model data) via DAVE and to allow the user exploiting the data with the maximum flexibility through the Jupyter notebook interface.

3.1. Eyjafjallajökull volcanic eruption (2010) with DAVE

The on-line coupled chemical transport model WRF-Chem was used to simulate the dispersion of the volcanic ash cloud emitted during the Eyjafjallajökull volcanic eruption in 2010. Comparisons of the predicted ash cloud

with satellite observations were used to examine if the model can predict the distribution of the ash cloud.

The following figures (Figs. 2 to 4) show the location of the volcanic plume which is represented by aerosol optical thickness (AOT) calculated from WRF-Chem simulations and the total ash load observed by the Spin Enhanced Visible and Infra-Red Instrument (SEVIRI). All these figures reveal that the model catches the observed ash load quite well. Note, however, that these different atmospheric parameters allow only a qualitative comparison. Both datasets were ingested into TAMP. The SEVIRI sensor operates on the Meteosat Second Generation (MSG) platform. MSG is a geosynchronous satellite, orbiting the Earth at a height of 35,800 km with a period of 24 hours and a nadir point at approximately 0° longitude over the equator. Currently one operational (Meteosat-10), one supplementary (Meteosat-9, 9.5°E) and one back-up (Meteosat-11, 3.4°W) European geostationary satellites are in orbit.

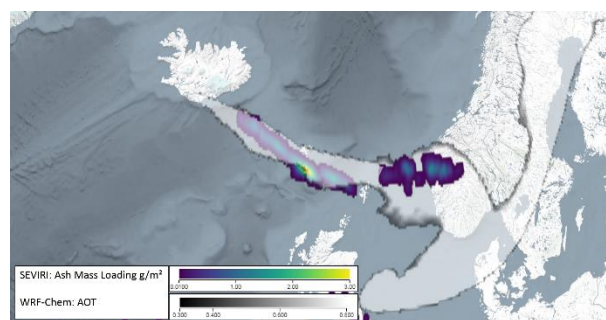


Fig. 2. Comparison of simulated AOT with total ash load from SEVIRI on 15th April 2010 15 UTC

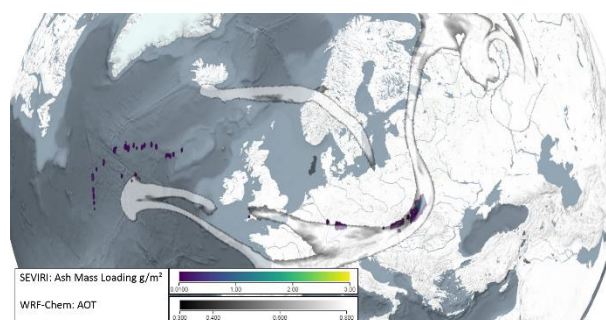


Fig. 3. Comparison of simulated AOT with total ash load from SEVIRI on 17th April 2010 17 UTC

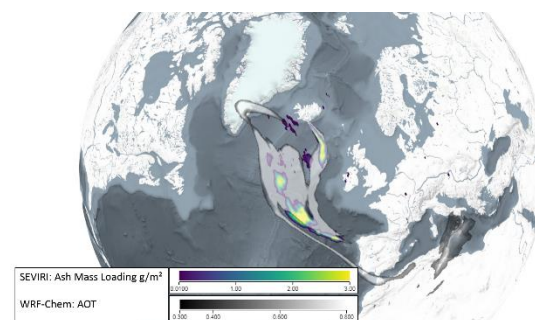


Fig. 4. Comparison of simulated AOT with total ash load from SEVIRI on 7th May 2010 18 UTC

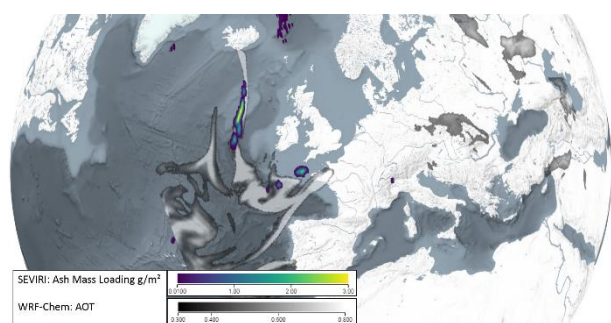


Fig. 5. Comparison of simulated AOT with total ash load from SEVIRI on 11th May 2010 21 UTC

3.2. Holuhraun volcanic eruption (2014) with Jupyter notebook

Different SO₂ concentration datasets related to the Holuhraun volcanic eruption (2014) have been loaded on TAMP: FLEXPART numerical simulations, OMI and GOME-2 satellite data, ground measurements from Austrian local authorities. All collections are available for exploitation through both DAVE and the Jupyter notebook user interfaces. While DAVE allows visually comparing time and space co-located data sources, by means of Jupyter it is possible to write down Python code to directly access all available data and perform customised operations (e.g., data extraction, processing, download). The OWSlib package [3] that supports WCS 2.0 is provided, allowing performing the three main WCS operations, namely getCapabilities (to list all collections available on the TAMP WCS server), describeCoverage (to extract available information for a specific collection), and getCoverage (to extract data from a specific collection within a geographic bounding box and a time interval).

To demonstrate the flexibility and the easiness to access and display data, a single query was performed to collect SO₂ total column data over three cities Edinburgh (Scotland), Vienna (Austria), and Munich (Germany) in a three and half days timeframe. The query has the following syntax:

```
coverage_file=my_wcs.getCoverage(identifier=['FLEXPART_SO2_2D_175_DU_176_4326_01'],
format='application/xml', subsets=[('Long',-3,-3),
('Lat',56,56),('t',1411171200,1411473600)] )
```

The query returns data from the FLEXPART SO₂ integrated column collection ('FLEXPART_SO2_2D_175_DU_176_4326_01'), in XML format, for the point Lat: 56°N, Lon: 3°W (Edinburgh), in the timeframe 1411171200,1411473600 (unix times corresponding to 2014-09-20 00:00:00 and 2014-09-23 12:00:00 respectively).

The same query is repeated two more times changing only the point coordinates for the other two cities. Figure 6 shows the resulting plot (displayed using matplotlib), where

it is easy to identify the occurrence and intensity of the volcanic ash plume for all three cities at different times.

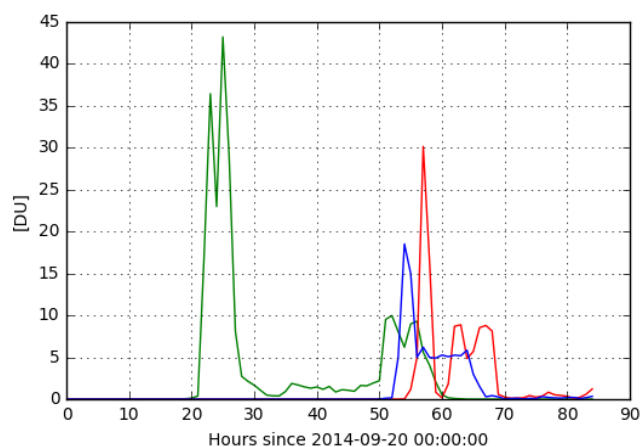


Fig. 6. Temporal evolution of FLEXPART SO₂ integrated column (in DU) over Edinburgh (green), Vienna (red), and Munich (blue).

4. CONCLUSIONS

The need of changing the data exploitation paradigm is becoming more and more urgent. This need is even more pressing for the atmospheric sciences community, due to the enormous amount of EO-based data that will become available in the next ten years with the launch of the atmospheric Sentinels (S5P, S4, S5), the ADM-Aeolus, and EarthCARE Earth explorer missions and the upcoming second generation of EUMETSAT polar orbiting satellites.

The VEEDAM project aims at demonstrating the feasibility of a VRE pre-operational environment. There is anyway the need of further developments to move toward an operational scenario, e.g., improving the harmonization among different data sources, allowing comparing / combining measurements of the same atmospheric field (e.g., SO₂ of the planetary boundary layer) from different platforms.

5. REFERENCES

- [1] TAMP landing page <http://vtpip.zamg.ac.at/> (visited on 17.10.2017)
- [2] TAMP introductory video <https://www.youtube.com/watch?v=xWiy8h1oXQY> (visited on 17.10.2017)
- [3] OWSlib documentation page <https://geopython.github.io/OWSLib/> (visited on 17.10.2017)

FORESTRY-TEP RESPONDS TO USER NEEDS FOR SENTINEL DATA VALUE ADDING IN CLOUD

Tuomas Häme¹, Renne Tergujeff¹, Yrjö Rauste¹, Clive Farquhar², Peter van Zetten², Philip Kershaw³, Arnaud de Groof⁴, Jarno Hämäläinen⁵, Joost van Bemmelen⁶, Frank Martin Seifert⁶

VTT Technical Research Centre of Finland Ltd (1), CGI IT UK (2), Science and Technology Facilities Council (STFC) (3), Spacebel s.a. (4), Arbonaut Oy Ltd (5), European Space Agency ESA ESRIN (6)

ABSTRACT

The response from the user community has shown a great need for the Forestry Thematic Exploitation Platform (Forestry-TEP) that entered into the pre-operational phase in summer 2017. Forestry TEP offers a one-stop-shop for forestry remote sensing services. It makes data value adding quicker and smoother than traditional approaches that the users have applied. The platform offers access to imagery, computing infrastructure, ready-made value adding services and an opportunity for users to develop and upload their own applications to the platform. It forms a worldwide marketing channel for commercial remote sensing services and a networking forum for users.

Ten thematic processing services are presently available on the Forestry TEP - from computing vegetation indices and mapping of land and forest cover to biomass and change estimation. In addition, a user can utilize all the features of the Sentinel Application Platform (SNAP) image processing application, the Monteverdi/Orfeo toolbox, and the open source QGIS tool. The platform will move to an operational phase in 2018.

Index Terms— forestry, earth observation, remote sensing, platform

1. INTRODUCTION

The Copernicus Program of the European Union is a significant contributor in the growth of large volumes of Earth Observation (EO) data. The European Space Agency (ESA) predicts that the amount of data from the Sentinel satellites of Copernicus and ESA heritage missions grows exponentially reaching 50 000 Terabytes by 2022 when the volume today is less than 20 000 Terabytes. The Sentinel-2 constellation with two satellites alone provides 1.6 terabytes compressed data daily [1].

The growing availability of big data enable data value adding services that have not been possible before but the vast data volumes also force developing novel approaches for the practical work. For instance, it is practical to avoid downloading large data volumes whenever possible. This changes the paradigm of downloading data and analyzing those using in-house computing facilities. The analysis is

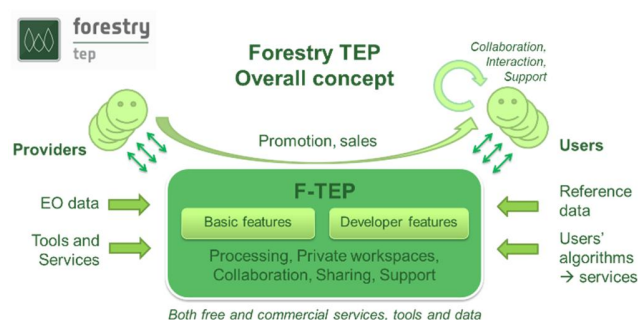


Figure 1. Forestry TEP overall concept: bringing together users from various segments with providers of data, tools and services.

brought close to the data and scalable cloud computing infrastructure is used in value adding. The users operate the computing facilities with software over the internet. The cloud based web services make it also much easier to share the analysis results and market them commercially worldwide. The Service Oriented Architecture (SOA) concept [2] that was developed more than ten year ago built a foundation to the present web-based cloud services.

This paper discusses building novel web services of large satellite data to the traditionally conservative forestry community using the Forestry Thematic Exploitation Platform or Forestry-TEP.

2. FEATURES OF FORESTRY TEP

2.1. General concept

Forestry-TEP is being built as a one-stop-shop for forestry remote sensing services (Figure 1). The platform includes the computing infrastructure, access to imagery, private workspaces with opportunity to share files openly or to the selected users, ready-made value adding services, and an opportunity for users to develop and upload their own applications. It also offers a worldwide marketing channel to commercial remote sensing services and a forum for user federation and networking.

In addition to the processing services available on the platform, a user can utilize all the features of the Sentinel Application Platform (SNAP) image processing application,

Table 1. Compatibility of Forestry TEP to requirements of Service Oriented Decision Support Systems. Requirements defined by Demirkan and Delen (2013).

Requirement for Service Oriented Decision Support System	Forestry TEP response
Accurate	Satellite and reference data are processed to value added information products to improve knowledge in decision making
Secure	Data safety and security functionality included
Governance	Governance system built
Compliance	European Cooperation Space Standardization, container-based technologies such as Docker™, Open Source Software, OGC Web Services standards adhered
Collaborative	Users are closely involved in the development and included in the consortium
Intra- and Inter-organizational	Developers and users represent versatile backgrounds
Synchronous	Data and process management system in place
Agile	New services can be implemented within days because of the modular structure
Commoditization	Open source software in central role
Adaptive	Agile service implementation with federation of service providers
Reuse and integration	Already the present services use same software components
Virtual	Cloud service using third party's infrastructure and software
Service-oriented	The user does not have to be aware of details of earth observation but can make advanced processing on the base of the needs.

the Monteverdi/Orfeo toolbox, and the open source Geographic Information System QGIS.

Forestry-TEP realizes concepts of Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). It fulfils general requirements for a Service-Oriented Decision Support System (Table 1).

The Forestry-TEP (forestry-tep.eo.esa.int/) is being developed by VTT Technical Research Centre of Finland Ltd as the coordinator and application and user specialist, CGIIT UK as the system developer and integrator, Science and Technology Facilities Council (STFC, UK) as the principal data access and infrastructure provider, and Spacebel (BE) and Arbonaut (FI) as application and service experts.

2.2. Data search and management

The user can search for data on the Forestry-TEP Platform, switch to access remotely any of the above-mentioned software tools, process data using these tools, save the result to the platform, and continue processing there or download the result to their own computer.

The platform presently offers access primarily to Sentinel-1, Sentinel-2 and Landsat-8 data. More data types, including the Landsat archive and commercial data, will be introduced according to user interest. Users can search over a geographic region for the available satellite images but also for existing products that have been computed and shared by other users.

Relevant results from data search can be saved to a data basket, where they can also be returned to for future processing in another session.

Applications and computed products can be shared with designated users or openly for everybody. In the operational

phase starting in 2018, there will be a possibility also to sell products and offer application software as a service on the platform. Already now, a high level of data privacy and security functionality is in place.

For the infrastructure and data access services, the platform is currently relying on the facilities provided by STFC via the Climate and Environmental Monitoring from Space (CEMS) and the Centre for Environmental Data Analysis (CEDA) with the UK Collaborative Ground Segment. A technical demonstration of integrating Forestry-TEP with the Earth Observation Innovative Platform Testbed Poland (IPT) facility has also been performed.

In future, the Copernicus Data and Information Access Services (DIAS) that are planned to become operational in 2018 will be strong candidates for the provision of the computing infrastructure (<http://copernicus.eu/news/upcoming-copernicus-data-and-information-access-services-dias>).

2.3. Processing services

Presently ten thematic processing services are available on the Forestry-TEP (Figure 2): for computing vegetation indices and for mapping of land and forest cover, biomass and change.

The user can upload their own data to train *e.g.* the land cover or biomass models for services that require reference data. Alternatively, they can interactively generate the reference data using the GIS software that has been integrated with the platform.

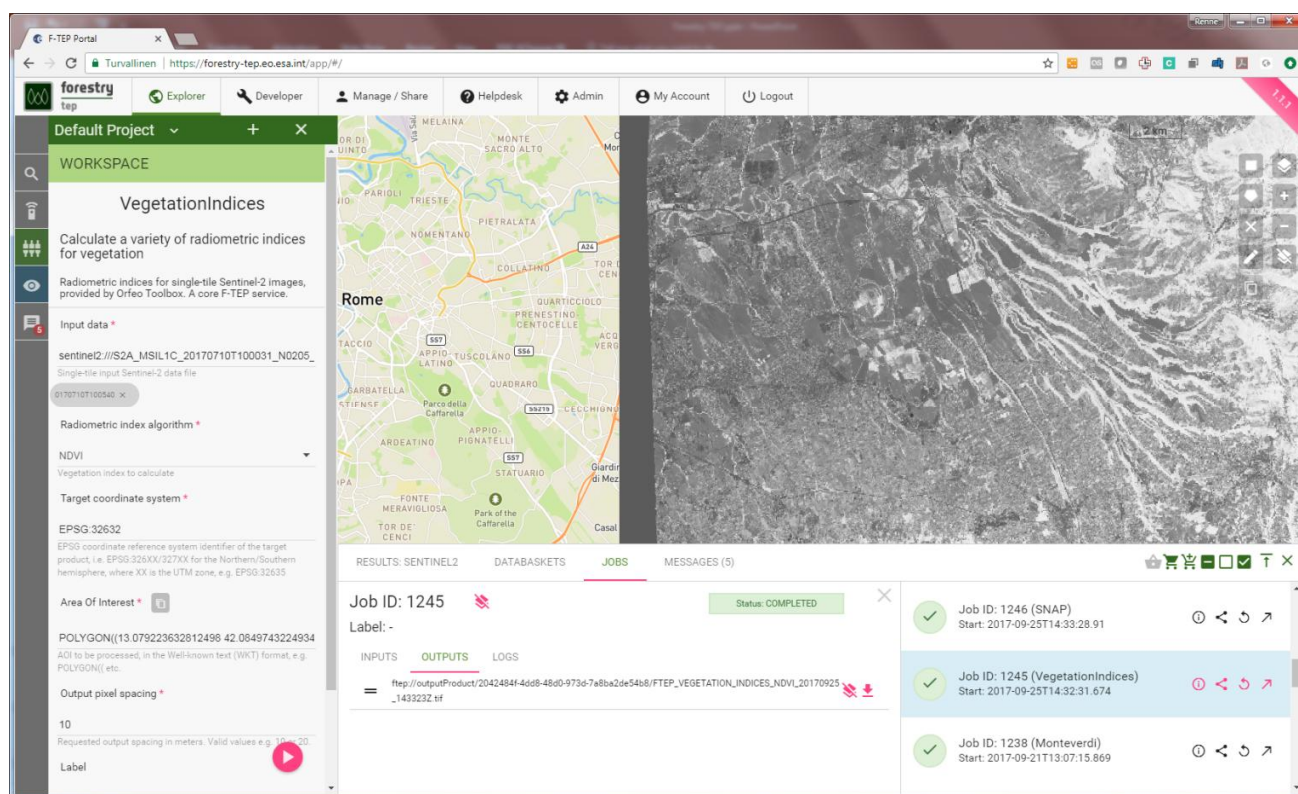


Figure 2. Forestry TEP platform user interface. On the left, parameter definition for the VegetationIndices service; on the bottom, details of the performed processing job; on top right, the output of the processing service, an NDVI vegetation index map.

Through a voting system available on the Forestry TEP web site, a user can affect the prioritization of future core services.

Service providers can publish their existing services or develop new services with the help of a web-based developer interface of the platform.

3. PILOTS

Two extensive pilot projects are currently being executed as part of the pre-operational phase of the development project.

One pilot service is related to helping the Reduction of Emissions from Deforestation and forest Degradation program (REDD), which is part of the Paris Agreement [4]. Forest cover is mapped in the states of Chiapas and Durango in Mexico with a total area of approximately 200,000 km². The mapping is done by the local users who are applying the platform supported by the Forestry TEP team.

The other pilot covering a land area of similar size is about mapping harmful broadleaved shrubs in forest regeneration areas of Finland. It is performed for the Finnish Forest Centre. The broadleaved shrubs cause significant damage and growth reduction to the conifer seedlings that are the primary species in forest management and lead to major economic losses. Measures for shrub removal are supported by the government, which leads to a responsibility for the authorities

to control the condition of the regeneration areas on the field. This causes substantial costs although only a fraction of regeneration areas can be field-checked. Satellite image analysis will reduce the costs for field checks while extending the coverage of law enforcement activities.

4. USER ACCEPTANCE

In September 2017, the number of registered users of Forestry TEP was 135 of which 50 were active during this month. Half of these active accounts were created in September. The growing user database includes members from academia, research organizations, private sector, public administration, and NGO's.

The users run several hundred jobs during the month. Four interactive on-line training sessions were given with 34 participants in total. Throughout the operation time of Forestry TEP, the image processing has focused on Sentinel-2 data. The proportion of processed Sentinel-1 images is approximately one tenth of the processed optical Sentinel-2 data.

During the initial community interviews in the Forestry TEP project, the major user interest was to have access to Sentinel data for download. When the users experienced the massive data sizes, their interest turned to cloud based processing services offered by the platform.

5. FUTURE OF THE PLATFORM

In the present project, the resources provided by ESA are relatively limited. Additional investments have to be made to build Forestry-TEP a full-scale operational service system and thus ensure European leadership in global value added service is earth observation. Preparations to transform without any interruption smoothly to the operational phase are ongoing.

The response from the user community has shown the great need for the Forestry-TEP. The users are eagerly expecting having a guarantee for the operational services of the platform to build their own services and commercial businesses on the platform.

6. REFERENCES

- [1] M. Drusch *et al.*, “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” *Remote Sens. Environ.*, vol. 120, pp. 25–36, 2012.
- [2] M. P. Papazoglou and W.-J. van den Heuvel, “Service oriented architectures: approaches, technologies and research issues,” *VLDB J.*, vol. 16, no. 3, pp. 389–415, Jul. 2007.
- [3] H. Demirkan and D. Delen, “Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud,” *Decis. Support Syst.*, vol. 55, no. 1, pp. 412–421, Apr. 2013.
- [4] United Nations, “Paris Agreement.” United Nations, pp. 1–25, 2015.

MONITORING URBANIZATION WITH BIG DATA FROM SPACE THE URBAN THEMATIC EXPLOITATION PLATFORM

Jakub Balhar^d, Thomas Esch^{a}, Hubert Asamer^a, Martin Boettcher^c, Enguerran Boissier^c, Andreas Hirner^a, Emmanuel Mathot^c, Mattia Marconcini^a, Annekatrin Metz^a, Hans Permana^b, Tomas Soukup^d, Soner Ureyen^a, Vaclav Svaton^e, Julian Zeidler^a*

^a German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen, Germany

^b Brockmann Consult GmbH, Geesthacht, Germany

^c Terradue Srl, Frascati, Italy

^d GISAT s.r.o., Prague, Czech Republic

^e IT4Innovations, VSB-Technical University of Ostrava, Ostrava-Poruba, Czech Republic

ABSTRACT

The capability to effectively and efficiently access, process, and analyze mass data streams from modern Earth observation missions such as the European Sentinels or the US Landsat program poses a key challenge. Hence, the implementation of operational, modular and highly automated processing chains, embedded in powerful hard- and software environments and linked with effective distribution functionalities, is of central importance. The TEP Urban platform aims at the utilization of modern information technology functionalities and services to bridge the gap between the existing mass data of the technology-driven Earth observation sector and the information needs of environmental science, planning, and policy related to the phenomenon of global urbanization. So far the TEP Urban system has successfully been used to process a variety of mass data collections of satellite imagery, including global coverages of more than 948,894 multispectral Landsat scenes, a global dataset of 25,550 Envisat-ASAR radar images, and several Sentinel-1 and Sentinel-2 data collections covering central Europe and Africa. The related TEP Urban services and products have yet been used by more than 240 institutions from 41 countries

Index Terms— Earth observation, Mass data, Urban, Exploitation platform, Open, Participatory

1. INTRODUCTION

A large-scale transformation has occurred on Earth, largely ignored in headline stories on the topic of global change – for some years now, the number of people living in urban areas has exceeded that of those living in rural regions. The trend towards urbanization shows no sign of abating. In particular, cities in Asia and Africa are expanding at a staggering pace.

Megacities are springing up in a matter of years and urban sprawls are emerging, spreading across extensive swathes of landscapes that, until recently, had been untouched by development or used for agricultural production. Today, approximately 7.2 billion people inhabit Earth. By 2050, this number will have risen to nine billion, 70 percent of which will be living in cities [1].

The global urbanization has regional roots, but it also comes with common drivers and causes. A global view is required to identify what they are. It is here that Earth observation (EO) can make a valuable contribution. It helps differentiate between urban and rural settlement forms and to introduce systems of categorization and delineation. Satellite-based geo-information delivers an up-to-date and comprehensive image of the built environment, while at the same time documenting its changes over time [2].

2. URBAN THEMATIC EXPLOITATION PLATFORM

The upcoming suite of Sentinel satellites in combination with their free and open access data policy will open new perspectives for establishing a spatially and temporally detailed monitoring of the built environment. However, the capability to effectively and efficiently access, process, analyze and distribute the mass data streams from the Sentinels - but also from other “big data” missions such as the Landsat program - poses a key challenge. This is also true with respect to the necessity of flexibly adapting the processing and analysis procedures to new or changing user requirements and technical developments. Hence, the implementation of operational, modular and highly automated processing chains, embedded in powerful hard- and software environments and linked with effective distribution functionalities, is of central importance.

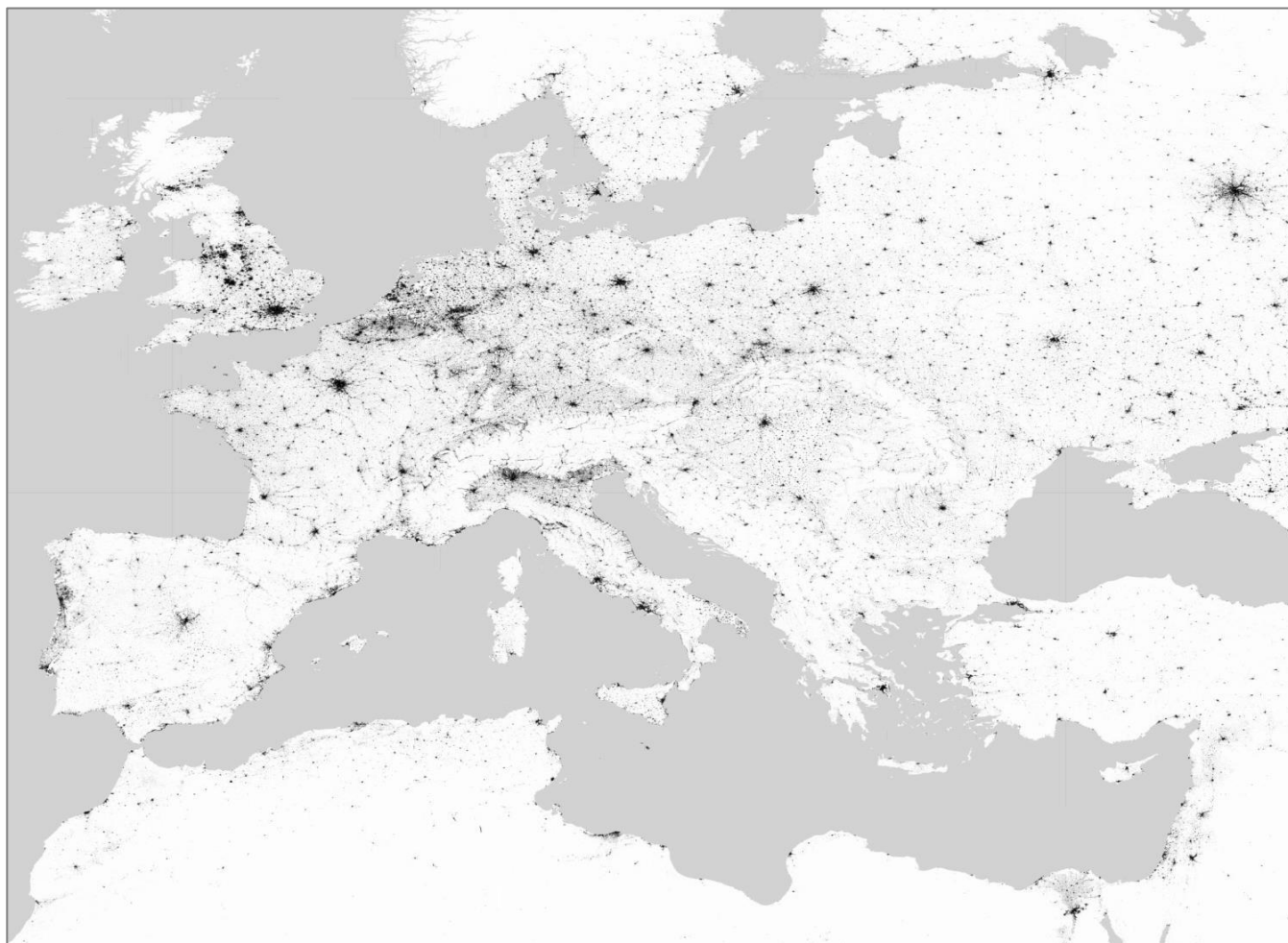


Fig. 1 GUF Image of Europe (Black = Urban, White = Non-Urban, Grey = Water bodies and No data)

Therefore, the TEP Urban platform [3] aims at the utilization of modern information technology functionalities (ICT) and services to bridge the gap between the technology-driven EO sector and the information needs of environmental science, planning, and policy. Key components of the system are an open, web-based portal connected to distributed high-level computing infrastructures and providing key functionalities for i) high-performance data access and processing, ii) modular and generic state-of-the-art pre-processing, analysis and visualization, iii) customized development and sharing of algorithms, products and services, and iv) networking and communication. These services and functionalities are supposed to enable any interested user to easily exploit and generate thematic information on the status and development of the built environment based on EO data and technologies.

The whole platform and all its parts are open source and its architecture allows adding new data sets and tools. Together, these functionalities and concepts support the four basic use scenarios of the U-TEP platform: (1) explore existing thematic content; (2) task individual on-demand analyses; (3) develop, deploy and offer your own content or

application; and (4) learn more about innovative data sets and methods.

3. RESULTS

So far, the TEP Urban system has successfully been used to process a variety of satellite mass data collections, including global coverages of multispectral Landsat scenes for the years 2015, 2010, 2000 and 1990 (948,894 images), a global dataset of 25,550 Envisat-ASAR radar images collected between 2010-2012, and Sentinel-1 and Sentinel-2 data collections covering several regions in central Europe and Africa. The results of this EO-based processing have been used to support the generation several new global thematic layers, including the Global Urban Footprint (GUF-2012/GUF+ 2015) binary settlement masks (Figure 1) [4], the TimeScan Landsat 2015/2010/2000 data sets, and the experimental GUF-DenS 2012 product derived from a combination of the GUF and TimeScan Landsat data.

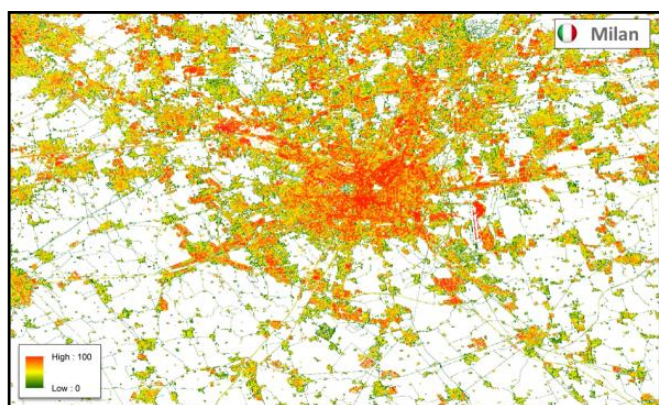


Fig. 2 GUF-DENS of Milan, Italy. It shows Urban density by means of Imperviousness/Greenness, ranging from low density (Green) to high density (Red).

The GUF-DenS shows the imperviousness (or as an inverse the greenness) of all settlements globally in 30m spatial resolution. Milan example is shown in Figure 2.

The TimeScan Landsat layers [5] represent higher-processing level baseline products derived from Landsat imagery that provides a cloud-free representation of the spectral and temporal characteristics of the land surface for the years 2015, 2010 and 2000 in a 30m resolution.

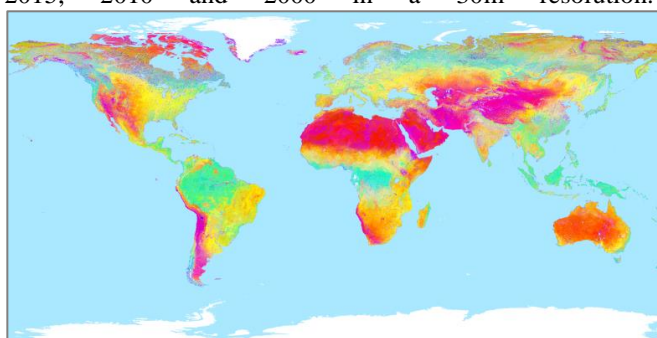


Fig. 3 Global TimeScan Landsat 2015 layer (30m ground resolution) derived from 452,799 Landsat-8 scenes with the Urban TEP System.

To prepare the products it was necessary to process more than 500TB of data. The product (Figure 3) shows the spectral and temporal characteristics of the land surface. It includes, for example, the mean and maximum NDVI and other indices over time. Color composites of these indices result in informative maps of urban areas. New York example is shown in Figure 4.

In addition, thematic processors were developed and deployed to generate TimeScan products from Sentinel-1, Sentinel-2 and Envisat-ASAR data. The algorithms are implemented in the processing centres of U-TEP. As a result, U-TEP can offer on-demand TimeScan data sets for selected regions. A more detailed specification and visualization of all products and services already available are provided at the Urban TEP website/portal [6] and in Table 1.

Table 1. Thematic layers and services of U-TEP

Layer/service	Number	Description
New global products	5	GUF2012, TimeScan Landsat 2015, GUF-DenS 2012, GUF-NetS 2012, GUF + 2015
New regional products	12	for six U-TEP demo cities and/or selected countries: GUF+ Evolution, TimeScan Sentinel-1/-2, LULC-Development, WorldPop-GUF, etc.
Services	4	TimeScan on-demand (Sentinel, Landsat), Functional Area Definition (urban-rural), Data Visualisation and Analytics.
Integrated global auxiliary data sets	7	Statistics, geo-data, thematic layers, social media a.o. WorldPop, Gridded Population World, Global Administrative Units, Nightlights, World Bank Statistics, UN Statistics.
Demonstrations	4	High Altitude Pseudo Satellites (HAPS), Monitoring of urban ground motion (PSI), Location-based services (e.g. tweets, traffic).

Combining prepared products such as GUF and TimeScan with auxiliary layers (e.g. population density, tweets density, nightlights) creates valuable and unique information for urban analysis. Example of a combination of Tweets and Nightlights in Figure 5. Also, the visualization plays a large role to make relevant urban characteristics visible. Such analysis is supported by the Visualization and Analytics toolbox implemented in U-TEP. It, of course, visualizes all data, layers and results, but it also allows to integrate own data (e.g., WMS, vector, raster, CSV). For customized areas and use cases, analysis can be performed and indicators and statistics based on provided and/or own data can be generated. It allows users to combine multiple

layers (EO, geo-data, statistics, and indicators) with

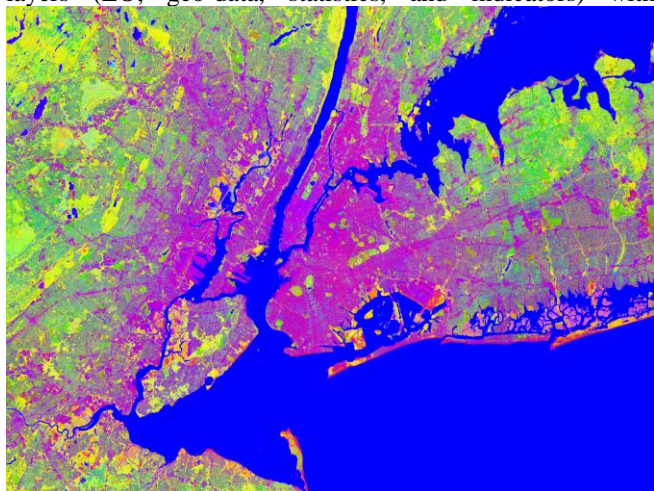


Fig. 4 TimeScan image of New York, USA. Color-composite of max. NDBI (Red), max. NDVI (Green) and mean NDWI (Blue).

choropleths and charts. Various types of charts can be created on the fly (e.g. column, pie, scatter, sortable tables). Finally, results can be shared via portal, URL or through screenshots or they can be exported in e.g. CSV format.

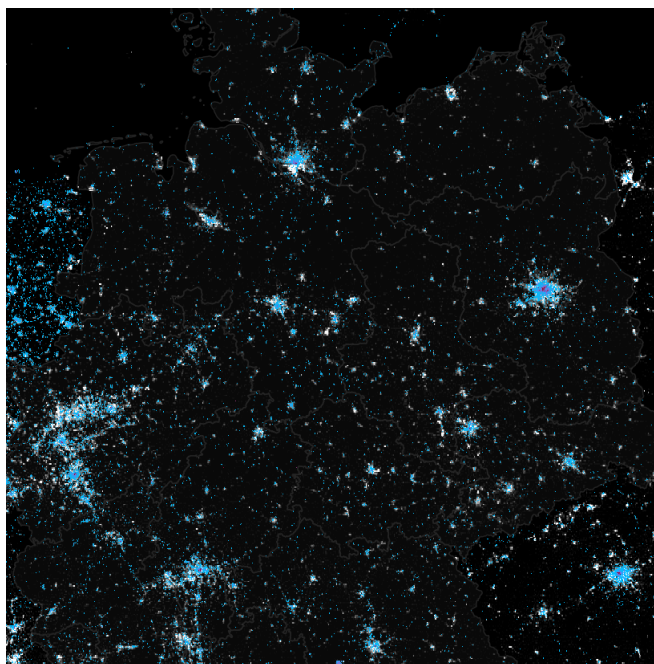


Fig. 5 Nightlights overlaid by the density of the tweets in Germany. The Nightlights are represented by the intensity of white while Twitter density is represented on the blue-red scale (red means higher density).

4. CONCLUSIONS

So far more than 240 institutions from 41 countries have requested TEP Urban data and system access. Therewith the Urban TEP platform is supposed to initiate a step change by providing an open and participatory platform that enables any interested user to easily exploit and generate thematic information on the status and development of the built environment from big data collections of satellite imagery – in particular Sentinel-1 and Sentinel-2, in combination with other sources such as geo-data, statistics and/or social media (e.g. geotagged Tweets).

5. REFERENCES

- [1] United Nations (2014). 2014 Revision of World Urbanization Prospects of the Population Division of the Department of Economic and Social Affairs of the United Nations. Available URL: <http://esa.un.org/unpd/wup/>.
- [2] Esch, T., Taubenboeck, H., Heldens, W., Thiel, M., Wurm, M., & Dech, S. (2010). Urban remote sensing e how can earth observation support the sustainable development of urban environments? In Proceedings of 46th ISOCARP Congress, 19-23 September 2010, Nairobi, Kenya.
- [3] Esch, T., Asamer, H., Boettcher, M., Brito, F., Hirner, A., Marconcini, M., Mathot, E., Metz, A., Permana, H., Soukop, T., Stanek, F., Kuchar, S., Zeidler, J., Balhar, J. (2016): Earth Observation-Supported Service Platform for the Development and Provision of Thematic Information on the Built Environment – the TEP-Urban Project. XXIII ISPRS Congress 2016, 12-19 July 2016, Prague, Czech Republic.
- [4] Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E. (2017): Breaking new ground in mapping human settlements from space. The Global Urban Footprint. ISPRS Journal of Photogrammetry and Remote Sensing. Submitted (Available URL: <http://arxiv.org/abs/1706.04862>).
- [5] Marconcini, M., Üreyen, S., Esch, T., Metz, A., & J. Zeidler (2017a): Mapping urban areas globally by jointly exploiting optical and radar imagery the GUF+2015 layer. ESA WorldCover 2017, 14-16 March 2017, Frascati, Italy. Available URL: <http://worldcover2017.esa.int/files/2.2-p2.pdf>.
- [6] European Space Agency (2017). Thematic Exploitation Platform, European Space Agency (ESA). Available URL: <https://tep.eo.esa.int/>.

FAST MI-SAFE PLATFORM: FORESHORE ASSESSMENT USING SPACE TECHNOLOGY

Joan Sala Calero¹, Gerrit Hendriksen¹, Jasper Dijkstra¹, Amrit Cado van der Lelij¹, Mindert de Vries¹, Rudie Ekkelenkamp¹, and Edward P. Morris²

¹ Deltares, Boussinesqweg. 1, 2629HV Delft, the Netherlands

² Universidad de Cádiz, 11510 Puerto Real, Spain

ABSTRACT

Foreshore Assessment using Space Technology [1] (FAST, 2014-2018, EU-FP7 607131) has created a platform for EU Earth Observation (EO) Programme Copernicus services, to support cost-effective, nature-based shoreline protection. Called “MI-SAFE”, it is developed with Open Source (OS) Intelligent Geographical Information System (IGIS) components from the OpenEarth [2] stack, and interoperable OGC standards. Processing large volumes of EO data for global coastal vegetation and elevation products is done using the cloud native Google Earth Engine platform. These are intelligently combined with other layers to form the inputs to XBeach [3], an OS hydrodynamic model developed by Deltares, UNESCO-IHE and TU Delft. Access to data is via a web-based user interface with different resolutions and complexity (Educational and Expert modes), and a Catalogue Service for the Web (CSW). MI-SAFE is intended to be a sustainable contribution to coastal Nature-based Solutions, increasing both awareness and facilitating advanced modelling, within the engineering community.

Index Terms— OGC protocols, Maritime safety, Marine Information System, Copernicus, Earth Observation, OpenEarth, XBeach, INSPIRE, Google Earth Engine, Coast, Nature

1. INTRODUCTION

Marine foreshores are currently not included in water safety assessments and levee design. However, foreshores deliver several services, such as increasing sedimentation, reducing erosion and attenuating waves that mitigate flood risk by improving levee stability and lifetime. Including foreshores in levee design and safety assessments can result in considerable cost reductions for flood risk management.

The FAST (Foreshore Assessment using Space Technology) project has developed a platform (MI-SAFE) to provide key data for modelling foreshores; such as elevation, morphology, sediment and vegetation properties. This includes new products derived from the USGS Landsat and Copernicus Sentinel EO missions, and field measurements, such as wave attenuation and erosion/deposition suitable for calibration/validation activities. Using these, the OS XBeach model was adjusted to include vegetation (XBeach-VEG) at eight characteristic case-study sites across Europe (Spain, Romania, United Kingdom, and the Netherlands).

Relationships between foreshore properties and wave attenuation were used for extensive XBeach-VEG simulations, which trained a Bayesian model; allowing rapid estimation of flood risk reduction for any observable combination of foreshore properties. User interaction with model results was facilitated by implementing a fully OS Intelligent GIS with a web viewer, extensive documentation, and a Catalogue Service for the Web (CSW). Results are presented in two modalities, focussing on 2 main user groups (Educational and Expert users).

2. EDUCATIONAL AND EXPERT MODES

The Educational mode gives a first indication of the presence, and potential flood risk reducing effects of foreshores. It uses a combination of available standard data products, such as global SRTM elevation, and new products, such as global intertidal elevation (section 4.1), and coastal vegetation presence (section 4.2), derived from processing large volumes of EO time-series data. Users can explore the contribution of vegetation to flood risk reduction, i.e., the results of the Bayesian model, at any one of 20000 profiles across the shorelines of the globe.

The Expert mode shows XBeach-VEG simulations made with high resolution data, and hydrodynamic boundary conditions for various future scenarios. At present this is mainly limited to the case-study sites; but should increase as more data is ingested. Providing a detailed indication of the effects of foreshores on wave attenuation, it also includes an example of a 2D XBeach-VEG simulation coupled with a flood model (LISFLOOD). It aims to show Expert users the potential of EO derived data products combined with OS modelling to contribute to the development of Nature-based Solutions for shoreline protection. Allowing engineers to optimally use the coastal foreshores existing ecological and landscape attributes to reduce costs.

The platform uses OS data structures, and OA data described following the INSPIRE metadata conventions. The web viewer is available at <http://fast.openeearth.eu>, data layers are accessible at <http://fast.openeearth.eu/geonetwork/>, and modelling support is available from the active XBeach community <https://oss.deltares.nl/web/xbeach>.

3. SYSTEM ARCHITECTURE

The multi-layered client-server architecture of the MI-SAFE platform can be subdivided in three main blocks (Presentation, Logic and Data) corresponding to a three-tier

system application (Figure 1). The first layer handles user interaction with the web platform (an instance of Delta Data Viewer, DDV), whereas logic is handled by PyWPS services in the back-end. XBeach is used for across shore wave modelling. Raster data is stored and accessed by GeoServer and vectors by PostgreSQL/PostGIS. GeoNetwork is used to edit and serve metadata. Last but not least, the cloud native Google Earth Engine platform is used to transform large volumes of EO data into global coastal data products.

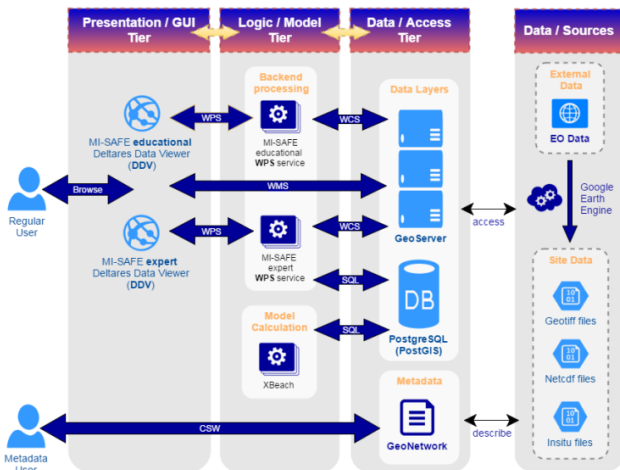


Figure 1. The multi-layered client-server architecture of the MI-SAFE platform.

3.1. Presentation Tier (Deltares Data Viewer)

The MI-SAFE platform uses a customized instance of the Deltares Data Viewer (DDV) based on OpenLayers, an OS GIS viewer from the OpenEarth stack (<https://github.com/openearth>). The viewer makes data requests via OGC WPS protocol to the back-end, which returns information to produce plots and reports. Visualization of data layers is via OGC WMS, with layer definitions harvested dynamically by querying the FAST GeoServer via GetCapabilities (WMS).

The structure of the viewer includes a canvas where maps are shown, a table of contents where layers can be toggled on or off, and a pop-up that provides visualization of the modelling output (Figure 2). User interaction includes selecting data layers, and clicking on a coastline to see model simulations. Coordinates of this point are sent to the Logic tier, and this initiates a search to find the nearest pre-defined coastline segment (within a buffer of 1 degree). Coordinates of the transect are sent back to the Presentation tier, and appear to the user as a 2 km long line perpendicular to the coast. Results for the transect are calculated in the Logic tier, returned, and graphically presented to the user as a 2D plot of the transect; designed to highlight vegetation presence, and potential contribution to wave attenuation. This includes contextual information such as the required

crest height with and without vegetation and attenuation coefficients.

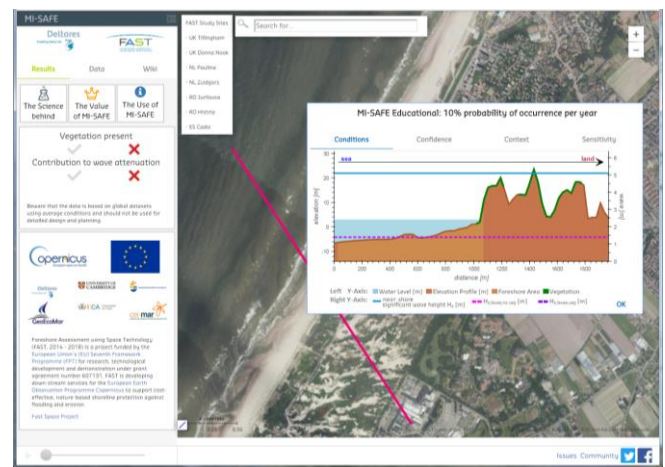


Figure 2. The DDV web GIS interface showing a transect result.

3.2. Logic Tier (PyWPS and XBeach)

PyWPS is used to get, calculate, and return transect results to the Presentation tier. This Python implementation of the OGC WPS protocol makes use of owslib and GDAL libraries to retrieve data values via WCS from the data layers in the Data Access tier.

The initial input of a clicked coordinate on the web map starts the search for a pre-calculated expert transect. In case no data is available an on-the-fly educational transect is computed (Figure 3). For the selected location, hydraulic boundary conditions in the form of off-shore wave data and surge levels are extracted from the ERA-interim dataset and the Global Tide and Surge Model respectively. Those are translated to onshore conditions on fixed return periods. For the educational mode, this is only a storm event return period of 10 years (an event with a likelihood of 10% in any given year), whereas for the expert mode 1% and 0.1% likelihoods are also shown in the plot.

Topography and vegetation properties (presence, type, and at study sites Leaf Area Index, LAI) of the transect are potentially obtained from the Data tier at different resolutions, and from several sources that may overlap. Hence, a selection rule is used that gives priority to pixels covered by the highest resolution layers (with the implicit assumption that they are better quality). Where EO derived LAI values are not available, i.e., the Educational mode, a standard LAI value is assigned to each pixel based on its type.

The matrix of variables (significant wave height, peak wave period, storm surge level, elevation, and LAI) returned from the Data tier can then be used to dynamically run XBeach-VEG simulation(s). However, depending on computing resources this may take some time, hence rapid access is provided by querying a pre-defined look-up table. This is derived from pre-calculated XBeach simulations (~

30 000) that represent the global range in variables, which were used to train a Bayesian model.

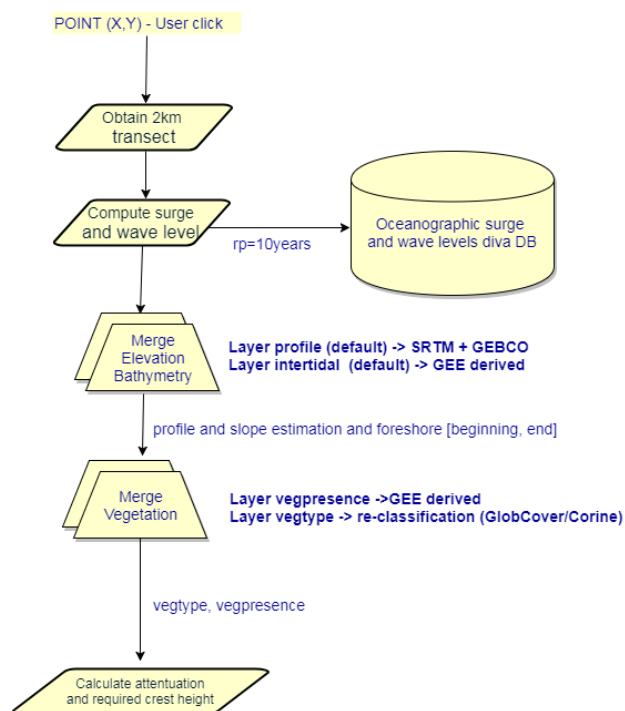


Figure 3. The MI-SAFE educational algorithm steps.

3.3. Data Tier (GeoNetwork, GeoServer, PostGIS)

For the Data tier OGC compliant solutions are used to handle raster and vector layers, and the corresponding metadata. GeoServer was chosen as the solution to handle global and local raster layers in GeoTiff format; providing WMS access for visualization, and WCS for coverage queries. GeoNetwork is deployed as a catalogue solution for metadata editing, and a discovery interface through the CSW protocol. Vector outputs produced by XBeach simulations are stored in a PostGIS database that enables fast queries from the PyWPS instance.

4. DATA LAYERS

MI-SAFE uses a mix of global, regional, and local data sets to provide both the educational and expert modes. These can easily be updated as improvements are released. Presently the standard global layers include; bathymetry (GEBCO), topography (SRTM), land cover and vegetation type (CORINE + GLOBCOVER reclassification), wave and surge information (diva world wave periods), and coastlines (OpenStreetMap). Regional and local datasets include; combined elevation/bathymetry for the Netherlands (AHN combined with vaklodigen), and high-resolution Digital Elevation Models (DEM) and LAI at the study sites.

Early on in the conception of the platform the difficulties of matching global elevation/bathymetry, poor coverage in intertidal zones, and low resolution of global land cover

maps, led us to develop two global datasets from high resolution EO data.

4.1. Global Intertidal Elevation

To improve the coverage in the gap between global bathymetry (GEBCO), and topography (SRTM) a global intertidal elevation map was derived using a combination of USGS Landsat and Copernicus Sentinel 2 images collected between 1997 and 2017.

Based on the traditional ‘waterline’ method [4, 5]; for each area-of-interest (AOI) on the coast, surface water was identified (using indices and classification) in a number of images (median of 317 images per AOI) with different tidal elevations. Composite, time-ensemble average (TEA) images of the probability of inundation were created, which were converted to intertidal elevation maps.

As assigning water levels to each image was not feasible at the global level, we developed a novel technique to transform TEA images of normalised difference spectral indices (NDSI) that represent water (here the NDSI of SWIR1 and Green bands), to elevation. Rather than segment every image into land-water, we normalised TEA-NDSI images by the spatially-averaged values of regions identified (using global elevation data sets) as land and water, respectively. This yielded a single image per AOI that represents inundation probability, and for each pixel in the intertidal zone, we assume that this represents the long-term average of tidal inundation.

As this inundation probability is derived from a collection of images that span a time period similar to the tidal epoch, i.e., the time period over which tidal height statistics are derived (commonly 19 years), then pixels with a probability of 1 represent permanent water, and have elevations less than or equal to the lowest astronomical tide (LAT), whereas land ($p = 0$) represents elevations more than or equal to the highest astronomical tide (HAT). By deduction, $p = 0.5$ is equivalent to local mean sea level (LMSL).

Global tidal statistics (LAT/HAT, 2005 to 2025) for each AOI were derived from the Global Tide and Surge Model (GTSM) [6], and used to rescale inundation probability images, giving an estimation of mean intertidal elevation (m, LMSL) in the period 1997 and 2017.

The process was carried out on ~ 20000 AOIs covering the global coast (defined using Open Street Map tiles) in Google Earth Engine, and took ~ 60 days. Validation of the predicted versus observed elevation at the case study sites, and other regions with quality intertidal elevation data suggested Root Mean Square Error (RMSE) values ranging between 0.3 and 1 m (Figure 3). Nevertheless, although better than the match between SRTM and GEBCO, systematic errors were observed in the product, related to the ability of the NDSI to define water, and availability of tidal statistics; suggesting there is room for improvement.

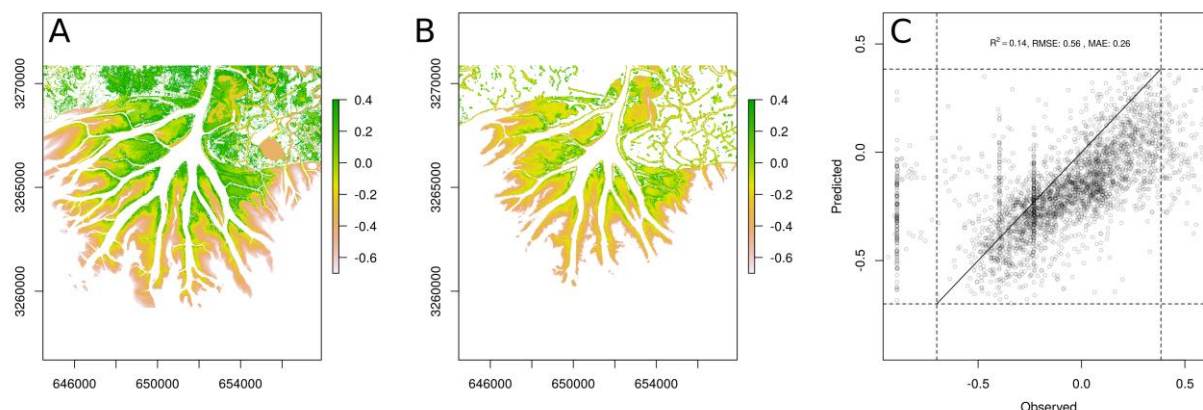


Figure 4. Example of validation of the FAST intertidal elevation product (m, LMSL) at Atchafalaya Delta, LA, USA. A) Observed elevation derived from the USGS Coastal National Elevation Database (CoNED) Project - Topobathymetric Digital Elevation Model (TBDEM, https://lta.cr.usgs.gov/coned_tbdem), B) predicted elevation, C) scatterplot of observed and predicted elevation.

4.2. Global Vegetation Presence

The low resolution of global land cover maps led us to develop a simple estimation the presence/absence of vegetation in coastal zones using USGS Landsat and Copernicus Sentinel 2 images between 2013 and 2017.

The binary presence of vegetation was determined by fitting a harmonic function to a time-series of Normalised Difference Vegetation Index (NDVI) images, and segmenting images representing the fitted harmonics using a threshold for both the mean, and amplitude of the NDVI harmonics. As for the intertidal elevation, processing was carried out on ~ 20000 AOIs covering the global coast (defined using Open Street Map tiles) in Google Earth Engine, and took ~ 20 days.

5. CONCLUSIONS

The MI-SAFE platform is an example of a marine IGIS assembled with OS components that provides a fully OGC compliant solution. The platform makes it easy to include new data, and can facilitate the demonstration of advanced commercially available data layers. One major advantage of the system is that processing of large volumes of EO data is carried out using the cloud native Google Earth Engine, which substantially reduces storage issues, and allows rapid development of new products.

The purpose of the platform is to support Nature-based Solutions for coastal defence, helping users understand how vegetated foreshores reduce coastal flood risk, and providing key data resources to enable expert users with their own modelling efforts. This knowledge will help to reduce the cost of flood protection, as well as deliver inputs towards wide-spread, successful restoration and conservation of coastal ecosystems.

6. REFERENCES

- [1] Morris, E. P., Jesus Gomez-Enri, and D. Van der Wal. 2015. 'Copernicus Downstream Service Supports Nature-Based Flood Defense: Use of Sentinel Earth Observation Satellites for Coastal Needs'. *Sea Technology* 56 (3): 23–26.
- [2] Van Koningsveld, M., G. J. De Boer, F. Baart, T. Damsma, C. Den Heijer, P. Van Geer, and B. De Sonnevile. 2010. 'OpenEarth-Inter-Company Management of: Data, Models, Tools & Knowledge'. In *Proceedings WODCON XIX Conference. Beijing, China*.
- [3] Van Rooijen, A. A., R. T. McCall, J. S. M. Van Thiel de Vries, A. R. Van Dongeren, AJHM Reniers, and J. A. Roelvink. 2016. 'Modeling the Effect of Wave-vegetation Interaction on Wave Setup'. *Journal of Geophysical Research: Oceans* 121 (6): 4341–59.
- [4] Mason, D. C., I. J. Davenport, and R. A. Flather. 1997. 'Interpolation of an Intertidal Digital Elevation Model from Heighted Shorelines: A Case Study in the Western Wash'. *Estuarine, Coastal and Shelf Science* 45 (5): 599–612.
- [5] Murray, Nicholas J, Robert S Clemens, Stuart R Phinn, Hugh P Possingham, and Richard A Fuller. 2014. 'Tracking the Rapid Loss of Tidal Wetlands in the Yellow Sea'. *Frontiers in Ecology and the Environment*, May. doi:10.1890/130260.
- [6] Muis, Sanne, Martin Verlaan, Hessel C. Winsemius, Jeroen CJH Aerts, and Philip J. Ward. 2016. 'A Global Reanalysis of Storm Surges and Extreme Sea Levels'. *Nature Communications* 7.

EFFICIENT AND LARGE-SCALE LAND COVER CLASSIFICATION USING MULTISCALE IMAGE ANALYSIS

François Merciol, Thibaud Balem and Sébastien Lefèvre

Université Bretagne Sud – IRISA
Campus de Tohannic, BP 573, 56017 Vannes Cedex, France

ABSTRACT

While popular solutions exist for land cover mapping, they become intractable when in a large-scale context (e.g. VHR mapping at the European scale). In this paper, we consider a popular classification scheme, namely combination of Differential Attribute Profiles and Random Forest. We then introduce new developments and optimizations to make it: i) computationally efficient; ii) memory efficient ; iii) accurate at a very large scale; and given its efficiency, iv) able to cope with strong differences in the observed landscapes through fast retraining. We illustrate the relevance of our proposal by reporting computing time obtained on a VHR image.

Index Terms— Big Data, Differential Attribute Profiles, Max-Tree, Land Cover Mapping, Large-Scale Classification

1. INTRODUCTION

With the proliferation of Earth Observation sensors, as well as their continuously increasing performances (spatial resolution, revisit time, etc.), remote sensing has entered in the Big Data era. While in this context, dedicated architectures (clouds, HPC) represent a major component and various experiments have been reported, there is still an effort to be made on the algorithms themselves to make them adapted to large-scale, data- and computationally-intensive challenges raised, such as sub-metric land cover mapping at a continental scale, as provided in some Copernicus products. Indeed, Europe with its area of 10 millions of sq.km corresponds to a map of 40 TeraPixels at 50cm (Pleiades resolution) or 100 TeraPixels at 31cm (WorldView-3 resolution).

In the context of land cover mapping, the standard approach is to first characterize each single pixel by some features extracted from the original image, and then apply a supervised classification technique. While various options exist for these two steps, we can still observe some trends in the remote sensing community. As far as feature extraction is concerned, beyond spectral information, multiscale spatial analysis has shown a strong ability to offer a discriminating characterization of various land use/land cover classes, with for instance the popular Differential Attribute Profile [1] and its many recent extensions (e.g., [2]). Supervised classification, where a model is first trained based on some reference data before being able to predict the class of a new pixel, has been achieved with many methods in remote sensing, the two most popular being Support Vector Machine (SVM) and Random Forest (RF). The latter has the ability to evaluate the relevance of the different dimensions of the feature space and their influence on the classification process. Besides, as a decision tree, it helps the understanding of the classification rules that are

used. It is thus a solution commonly adopted in remote sensing [3]. Furthermore, when combined with DAP, RF usually achieves better results. Thus, in the sequel of this paper, we will use such a combination “DAP+RF” as a baseline. For the sake of research reproducibility, we rely here solely on open-source solutions. RF implementation is provided by the Shark library¹, while DAP relies on our own implementation in the Triskele library² acting as a new remote module for the OTB framework³. Nevertheless, a significant gain in terms of classification accuracy has been achieved with deep learning [4, 5], and thus deep architectures are gaining increasing interest. However, these solutions still require a high computational cost and a heavy training process, and thus cannot be considered as a relevant solution for large-scale mapping yet.

We propose here an overall process that fits the large-scale requirements, minimizing the computational cost as well as the memory footprint. Furthermore, thanks to this efficiency, we are able to retrain a classifier for each novel scene to be analyzed, leading then to a straightforward approach to ensure robustness to the high variability of the land cover classes observed in the various acquired scenes. Beyond the overall process, we specifically focus on the feature extraction step for which we introduce a novel algorithm for efficient (both in time and space) tree construction that improves our former findings [6]. We also extend previous work on computing DAP on derived features such as NDVI [7], and demonstrate here the relevance of computing DAP on textural features.

2. OVERALL WORKFLOW

As already stated, large-scale classification brings three complementary issues that are tackled in this paper: i) computational complexity that is addressed through efficient algorithms and reduction of the data to be processed at the different steps; ii) memory cost that is addressed through optimizations allowing to limiting the memory footprint of the overall process; and iii) variability of the land cover class (e.g. spectral signature of the forest might differ between Mediterranean area and Scandinavia). In order to address these challenges, we rely solely on method efficiency. More precisely, we do not consider domain adaptation techniques to adapt the classification model to each novel scene to be mapped. We rather assume that, given a near real time overall process, a user is able to provide some samples, train a classification model from these samples, predict the labels for the unlabeled pixels, visually or quantitatively assess the accuracy of the produced map, and update the samples (and then the model) until the obtained accuracy is satisfying. Let us note that this process actually fits many operational contexts, where the accuracy

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-13-JS02-0005-01 (Asterix project); and SIRS for providing the use case, data, and funding.

¹<http://image.diku.dk/shark>

²<https://sourcesup.renater.fr/triskele>

³<https://www.orfeo-toolbox.org>

requirements impose some manual assessment/correction as a post-processing step. Our workflow is given in Fig. 1 and applied for each new scene to be mapped. To ensure efficiency in a large-scale context, parallelism is mandatory and has been made explicit.

1. add to the original image some additional bands (e.g., NDVI, texture, etc.); each novel band is built in parallel;
2. compute a min- and/or a max-tree per band; subtrees are built in parallel for each tile, and then merged together; for the sake of parallelism, the tiles contain a similar amount of pixels, but there is no restriction regarding their specific shape (see Sec. 3);
3. provide reference samples from existing maps, ground truth data, or visual analysis;
4. characterize samples, based on the two following steps:
 - (a) characterize each node by some attributes (e.g., area, standard deviation, moment of inertia, etc.); with attributes being incrementally computed from leaves to root, the parallelism occurs over nodes within each tree level;
 - (b) filter the tree and produce the full DAP feature vector; each feature vector (related to a unique pixel) is computed in parallel;
5. train the classifier on all features and evaluate it;
6. add new samples (step 3), characterize them (step 4) and update the model (step 5) as long as it is necessary; select the subset of relevant features for classification;
7. using the selected set of features, perform land cover mapping, i.e.:
 - (a) characterize all pixels;
 - (b) predict the classes using the trained model;
8. achieve a manual post-processing.

Fig. 1. Overall workflow

Classification is achieved with Random Forest (RF). Since RF is able to identify the important features in a classification process, we apply it on the full set of DAP features but only to the samples in steps 4 and 5; while in steps 7 (a) and (b), we consider all pixels but only a subset of the features. This actually leads to significant reduction of both computational and memory costs.

Feature extraction is performed with Attribute Profiles (AP), obtained by applying filters with increasing level on an input image, and efficiently computed from tree-based image representations. The original AP can be replaced by its differential version where differences between the filtered images form the feature vectors. Furthermore, instead of using the original image band, it has been shown recently that AP (or DAP) can be computed on derived features such as NDVI [7]. We follow this approach here and we propose to compute DAP over original bands, NDVI, as well as some additional bands bringing texture information. More precisely, we consider the L_1 norm of the image gradient computed with the Sobel masks. While there exists other popular texture descriptors (such as Haralick features), our choice has been motivated by the very low computational cost of Sobel gradient (i.e. each pixel is only

read 4 times). Let us note that the texture can then be computed over any of the input bands, including also the NDVI band. Computing DAP on Sobel information offers an efficient way to characterize the behavior of the edge information at multiple sizes.

3. TREE CONSTRUCTION

Our proposal assumes that the image features can be extracted very efficiently. The tree construction step is thus a key step in the process. In our previous work [6], we have introduced a novel algorithm for tree construction. We have however observed that, while the algorithm was intrinsically multi-threaded, the merging step might be particularly costly at the lowest levels of the tree.

We are introducing here a novel algorithm that avoids this shortcoming. It relies on the counting sort algorithm and modifies the underlying Tarjan's Union-Find algorithm (Alg. 1) to store additional information. We recall that this algorithm aims to build a tree structure. Since the tree construction is not a predictive process, some pixels might be put apart before realizing they form the same set. Merging the related subsets is then costly. In order to minimize such a cost, some nodes are identified as having more weight (higher score or rank) than others, and will be selected when two sets are to be merged. Sibling nodes (waiting for being merged) can also be encountered but the algorithm minimizes the size of siblings that have to be later integrated in the eldest. The tree structure is stored in an array called *parents*, while the array *rank* contains scores allowing to choose among the siblings. A fusion step remains mandatory.

In our modified algorithm (Alg. 2, with leader and count used instead of parent and rank in Tarjan's algorithm), we store the number of direct children of a node (set to 1 by default) instead of the so-called rank. When two nodes are linked, either they have the same value and then the one with more children is promoted as leader, or they have different values and then the one with the head value is the parent.

Based on this modified algorithm, we then derive the overall tree construction algorithm (Alg. 3). It relies on a more compact data structure, and involves a new merging strategy. The various optimizations lead to a linear complexity and a memory footprint divided by two. It is a fully parallel algorithm, with only the reindexing step requires one image scan and the merging achieved over the tile edges. Initialization (INITBUILDTREE) consists in setting all cells of the parent array to the maximal value. We will later be able to determine if a cell already knows its parent or not. The main construction algorithm (BUILDTREE) relies on a scheduler that will first divide the image surface in as many parts (called *tiles*) as available processors *nbCores*. Each tile is then analyzed in parallel to build the related subtree (BUILDSUBTREE). To do so, edges linking neighboring pixels are sorted in ascending order (given a specific metric, related to the kind of tree: min-tree, max-tree), considering the efficient counting sort algorithm. Each processor then uses these sorted edges to update *leader* and *count* using our modified Union-Find algorithm. When linking two pixels (LINKLEADER), it becomes easy to reach the two roots. Only them are impacted by the change of children and juniors. Through this scan, we also update the direct link to the highest parent at each analysis of a branch. A new node *r* is created during the first link with a leader pixel (CREATECOMP).

Once partial trees have been computed by each processor, a merging step has to be performed. Let us underline that it may lead to a topology change in the tree, rendering this process particularly costly (see [6]). We thus optimize it by sorting all edges linking border pixels in a row (second loop in BUILDTREE). The merging


```

procedure INITUNIONFIND
  foreach pixel  $p \in I$  do
     $parent(p) \leftarrow p$ 
     $rank(p) \leftarrow 0$ 

function FINDROOT(pixel  $p$ )
  if  $parent(p) = p$  then
    return  $p$ 
   $parent(p) \leftarrow FindRoot(parent(p))$ 
  return  $parent(p)$ 

procedure UNION( $a, b$ )
   $pa \leftarrow FindRoot(a)$ 
   $pb \leftarrow FindRoot(b)$ 
  if  $pa \neq pb$  then
    if  $rank(pa) < rank(pb)$  then
       $swap(pa, pb)$ 
     $parent(pb) \leftarrow pa$ 
     $rank(pa) \leftarrow rank(pa) + 1$ 

```

Algorithm 1: Original Tarjan’s Union-Find algorithm

is applied on these edges (MERGEANDCOMPRESS), through a zipping that can lead to a topology change in the tree, thus justifying why we are processing smallest edges first. We then assign a new rank for each node (processing all siblings in a row while reaching the eldest). We finally scan the set of nodes and set them at the appropriate location.

We finally add reverse links from the parents to the children (LINEARBUILDCHILDREN). We thus follow a strategy similar to the counting sort. We compute a cumulative sum of children counts and use the underlying array to indicate the first available location and set the children in the children array subsequently.

4. EXPERIMENTS

We report here some experiments conducted in order to assess the performance of the proposed workflow, including the novel tree construction algorithm. The underlying architecture is a computation node with 2 sockets L5640 2.27 GHz, 24 dual-cores and 40 GB of RAM. The input data mostly consist of pansharpened VHR optical images coming with a 16-bit resolution.

We consider here an excerpt of a WorldView-3 color (RGB) image, of size $9,250 \times 10,408$, i.e. ca. 100 millions of pixels. We append to the original spectral bands some derived features, NDVI and some Sobel indices. Then, a min-tree and/or a max-tree is built from each selected band before computing DAP using some predefined thresholds. The feature extraction scheme leads to a feature vector of typical length varying from a few tens to more than one hundred (e.g., computing both a min- and max-tree on the 3 original bands as well as 2 derived bands, and considering 12 thresholds with a single attribute, leads to 120 features per pixel).

In this example, we have added NDVI as an additional band, as well as the Sobel gradient from NDVI band. We have then computed a max-tree only on the NDVI and Sobel bands. The number of thresholds has been limited to 4, thus leading for each pixel to a feature vector of 13 (4 attribute values for each tree and the 5 input bands: RGB, NDVI and Sobel on NDVI). As such, the memory footprint of the features has been reduced by a factor of 9, from 21.5 GB to 2.3 GB storage. As far as the CPU time is concerned, we refer the reader to Tab. 4. We can see that the tree construction step is less than 10 seconds per tree (or per band), and the overall

```

procedure INITLEADER
  foreach pixel  $p \in I$  do
     $leader(p) \leftarrow \infty$ 
     $count(p) \leftarrow 1$ 

procedure UPDATELEADER(pixel  $min$ , pixel  $max$ )
  while  $min \neq max$  do
     $up \leftarrow leader(p)$ 
     $leader(p) \leftarrow l$ 
     $p \leftarrow up$ 

function FINDUPDATELEADER(pixel  $p$ )
   $l \leftarrow p$ 
   $up \leftarrow leader(l)$ 
  while  $up \neq \infty$  do
     $l \leftarrow up$ 
     $up \leftarrow leader(l)$ 
   $UpdateLeader(p, l)$ 
  return  $l$ 

function LINKLEADER(pixel  $max$ , pixel  $min$ , bool  $eq$ , pixel  $a$ , pixel  $b$ )
  //  $max, min$  are leader pixels of  $a, b$ 
  //  $max$  is the parent with highest weight
  //  $eq$  is true if both weights are equal
  if  $eq$  AND  $count(max) < count(min)$  then
     $swap(max, min)$ 
     $count(max) \leftarrow eq ? count(min) : 1$ 
     $UpdateLeader(a, max)$ 
     $UpdateLeader(b, max)$ 
  return  $max$ 

```

Algorithm 2: Proposed adaptation of Union-Find algorithm

feature extraction step is achieved in about 20 seconds. The training achieved by the Random Forest classifier is performed in less than 5 seconds, considering a set of 15,000 training samples. With such a low computational cost for feature extraction and training, it is possible to rerun these steps (choosing other bands, attributes or thresholds, providing other training samples) until reaching a satisfying classification model. Finally, the prediction is done in about 1.5 minute. When including the other steps not given here for the sake of concision (e.g., I/O and memory ops), the land cover map is produced in less than 3 minutes. Let us note that with a standard, single-threaded implementation of the tree-related algorithms, the CPU time would have grown from about 20 seconds to possibly more than 15 minutes (i.e. more than 40x).

Step	Total	#	Min	Max
Tree construction	17.2"	2	8.4"	8.8"
Feature embedding	2.4"	2	1.2"	1.2"
Tree filtering	1.9"	2	0.9"	1.0"
Training	4.7"	1	–	–
Prediction	1'30.9"	1	–	–
Total	2'21.0"	–	–	–

Table 1. CPU cost evaluation: total time of each step, number of processes, and min/max CPU times (– when not applicable).

5. CONCLUSION

In this paper, we have addressed the land cover mapping problem at very large scale (e.g., paneuropean). To do so, we have considered as a baseline the popular scheme consisting of DAP feature extraction followed by classification using Random Forest. We have then focused on the feature extraction scheme and introduce novel algorithms to lower both the memory footprint and the computational cost, making the proposed implementation compatible with multi-threaded environments. The overall processing chain has been validated by SIRS in an operational context, namely through the mapping of Small woody features (SWF) for EEA39, and is part of the Triskele library, a remote module of the Orfeo ToolBox (OTB) CNES Open Source suite. Let us note that the proposed algorithm aims to be run on a server side, while client access to tree based-structures has been proven to be an effective solution [8].

Future work will include integrating recent DAP extensions (e.g. [2]) in order to improve the classification accuracy, as well as further investigating computation/memory cost optimization, and experimentally comparing the proposed implementation with existing ones [9, 10].

6. REFERENCES

- [1] M. Dalla Mura, J.A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3747–3762, 2010.
- [2] M.T. Pham, S. Lefèvre, and E. Aptoula, "Local feature-based attribute profiles for optical remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2017, to appear.
- [3] M. Belgiu and L. Drăguț, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.
- [4] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multi-modal and multi-scale deep networks," in *Asian Conference on Computer Vision*, 2016.
- [5] N. Audebert, B. Le Saux, and S. Lefèvre, "Joint learning from earth observation and openstreetmap data to get faster better semantic maps," in *IEEE/ISPRS Workshop on Large Scale Computer Vision for Remote Sensing Imagery*, 2017.
- [6] J. Havel, F. Merciol, and S. Lefèvre, "Efficient tree construction for multiscale image representation and processing," *Journal of Real-Time Image Processing*, pp. 1–18, 2016.
- [7] B.B. Damodaran, J. Höhle, and S. Lefèvre, "Attribute profiles on derived features for urban land cover classification," *Photogrammetric Engineering and Remote Sensing*, vol. 83, no. 3, pp. 183–193, 2017.
- [8] F. Merciol, A. Sauray, and S. Lefèvre, "Interoperability of multiscale visual representations for satellite image big data," in *ESA Conference on Big Data from Space*, 2016.
- [9] E. Carlinet and T. Géraud, "A comparative review of component tree computation algorithms," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3885–3895, 2014.
- [10] J. Kazemier, G. Ouzounis, and M. Wilkinson, "Connected morphological attribute filters on distributed memory parallel machines," in *International Symposium on Mathematical Morphology*, 2017, pp. 357–368.

```

procedure INITBUILDTREE(Image I)
  foreach pixel  $p \in I$  do
    parent( $p$ )  $\leftarrow \infty$ 

procedure BUILDTREE(Image I)
  tiles  $\leftarrow$  tileImage(I, nbCores)
  foreach tile  $\in$  tiles do in parallel
    BuildSubTree(tile, I)
  borders  $\leftarrow$  borderImage(I, tiles)
  foreach border  $\in$  borders do in parallel
    borders(border)  $\leftarrow$  getSortedEdges(I, border)
    MergeAndCompress(borders)
    LinearBuildChildren()

procedure BUILDSUBTREE(tile, I)
  edges  $\leftarrow$  getSortedEdges(I, tile)
  foreach  $e \in$  edges do
    ra  $\leftarrow$  FindUpdateLeader(e.a)
    rb  $\leftarrow$  FindUpdateLeader(e.b)
    if ra = rb then
      continue
    wa  $\leftarrow$  value(ra)
    wb  $\leftarrow$  value(rb)
    if wa < wb then
      swap(ra, rb)
    r  $\leftarrow$  LinkLeader(ra, rb, wa == wb, la, lb)
    par  $\leftarrow$  CreateComp(r, value(e), count(r))
    if ra  $\neq$  r then
      linkParent(ra, par)
    if rb  $\neq$  r then
      linkParent(rb, par)

function CREATECOMP(leader, level, size)
  par  $\leftarrow$  parent(leader)
  if par =  $\infty$  then
    par  $\leftarrow$  nbComp++
    parent(leader)  $\leftarrow$  par
    compLevel(par)  $\leftarrow$  level
    nbChild(par)  $\leftarrow$  size
  return par

procedure MERGEANDCOMPRESS(borders)
  foreach  $e \in$  borders do
    connectLeaf(e.a, e.b, value(e))
  newRank  $\leftarrow$  0
  foreach  $e \in$  tiles do
    findUpdateTop(e)
    parNewPos(e)  $\leftarrow$  newRank++
  foreach pos  $\in$  comp do
    swap(parent(pos), parent(parNewPos))

procedure LINEARBUILDCHILDREN()
  sum  $\leftarrow$  0
  foreach comp do
    sum  $\leftarrow$  sum + nbChild(comp)
    nbChild(comp)  $\leftarrow$  sum
    posChild  $\leftarrow$  nbChild
  foreach  $x \in$  leaf and comp do
    par  $\leftarrow$  parent(x) - 1
    pos  $\leftarrow$  posChild(par)
    pos++
    posChild(par)  $\leftarrow$  pos
    children(pos)  $\leftarrow$  x
    // first child : children(nbChild(par - 1))
    // last child : children(nbChild(par))

```

Algorithm 3: Proposed algorithm for max-tree construction

UNSUPERVISED OBJECT DETECTION ON REMOTE SENSING IMAGERY USING HIERARCHICAL IMAGE REPRESENTATIONS AND DEEP LEARNING

Nikki Aldeborgh, Georgios K. Ouzounis and Kostas Stamatiou

DigitalGlobe Inc.
1300 W. 120th Ave Westminster, CO 80234, USA

ABSTRACT

A new paradigm for large-scale, unsupervised object detection on remote sensing imagery is proposed, which relies on the synergy of hierarchical image representations and deep learning. The proposed paradigm: (a) reduces the search space to a set of candidate objects which conform to the geometric characteristics of the object of interest, hence dramatically decreases deployment time in comparison to brute-force approaches which scan the entire image; (b) discards the need for manual training data generation which is laborious, expensive and prone to user bias. An example application is presented where the max-tree and the VGG-16 convolutional neural network architecture are used for the detection of circular tanks on very high resolution satellite imagery.

Index Terms— Hierarchical image representation, deep learning, max tree, convolutional neural network.

1. INTRODUCTION

One of the main drivers of satellite imagery analytics is the rapid and accurate detection of objects of interest over broad areas. Deep learning has been successfully applied to image classification and object detection on multimedia imagery, achieving remarkable performance on benchmark data sets [1, 2, 3]. However, object detection on satellite imagery using deep learning entails unique challenges related to scale, such as: (a) The procurement of high-quality training data, in the form of examples of the objects of interest in a variety of settings; this usually involves the exploration of large satellite images to identify and outline said objects, which is time consuming and prone to error, given that the object size is usually small and that the resolution of the imagery obtained from current state-of-the-art sensors is at most 30cm/pixel. (b) Computational complexity resulting from the size of the search area which is typically very large; a brute-force approach is to slide a small window across the entire area and detect the objects within the window, which quickly becomes computationally prohibitive as the window size and/or the step size decrease.

The author names are listed alphabetically.

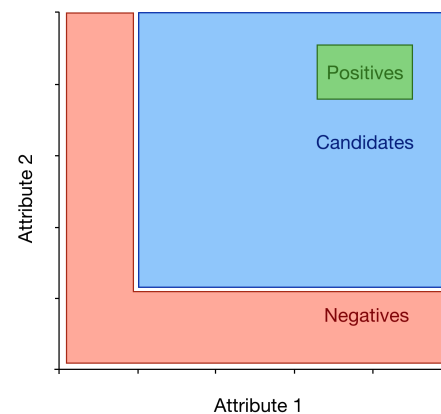


Fig. 1. Candidates, positives and negatives in a two-dimensional attribute space.

Our goal is to harness the discriminative power of deep learning while removing barriers for applying it successfully at the scale typically required by remote sensing applications. The premise of the proposed approach is that the geometrical characteristics of the object of interest should be used to *reduce the search area* and to *automate or, at the very least, facilitate the generation of training data*. In particular, hierarchical image representation structures such as the max tree [4] organize the image information content in connected pixel components which can be attributed with a plurality of metrics including size, compactness, elongation, contour smoothness etc. Provided that the object geometry can be described by a set of attribute constraints referred to as the *search space*, three groups of components can be extracted from the tree based on the attribute values, as shown in Fig.1: (1) in the search space; call these *candidates* as they include the object of interest, as well as other objects with similar geometry; (2) ‘well within’ the search space; call these *positives* as they most likely only include the object of interest; (3) ‘outside’ the search space; call these *negatives* as they most likely don’t include the object of interest.

The main idea is to train a neural-network-based classifier using the positives and negatives, and then deploy the classifier on the candidates. It is postulated that the neural

network can *learn the defining features of the object of interest from the positive and negative examples generated on a purely geometrical basis*, and, with this knowledge, can successfully discriminate the object of interest from irrelevant objects in the candidates. The approach obviates the need for manually generated training data, thus reducing training time, and narrows down the search area, thus reducing deployment time. In addition, using attributes which are scale-, rotation-, translation- and intensity-invariant, it can be applied across different scenes, geographies and sensors.

The remainder of the paper is devoted to the application of the proposed framework for the detection of circular tanks on very high resolution satellite imagery. The selected use case serves to illustrate the applicability of the framework, which is otherwise general and can be applied to other objects of interest, provided that they can be adequately described by certain geometrical attributes. Section 2 introduces some basic notions of hierarchical image representation structures. Section 3 describes the process of automatically and semi-automatically generating positive and negative examples using the max-tree, and training a VGG-16 neural network [5]. Section 4 reports and comments on accuracy metrics in select locations.

2. HIERARCHICAL IMAGE REPRESENTATIONS

Hierarchical image representation structures are dendrograms that organize the image information content in a manner that is both structured and ordered. Structure is defined by set-connectivity rules that dictate which pixels are clustered into groupings, referred to as *connected components*, and how.

In the case of the max-tree [4], pixel connectivity may adhere to the standard notion of adjacency-constrained connections [6], or to spatial generalizations of the notion of connectivity [7, 8]. Connected components are ordered with respect to intensity; bright components are nested within larger, darker components. In the case of the min-tree, the nesting order is inverse, i.e., small dark components are nested within larger, brighter ones. Equivalently, the min-tree is the max-tree of the inverted image.

Any one connected component maps to a unique node and every node points to its parent in a uni-directed relation. The finest components define the leaves of the tree, while the coarsest component, referred to as the background component, defines the root node, which points to itself and coincides with the image definition domain. The path from a component to the root via all of its supersets is referred to as a root path.

3. TRAINING

Twenty collections from WorldView-2 and WorldView-3 that contained circular tanks were identified in the United States, Central and South America, Europe, the Middle East and

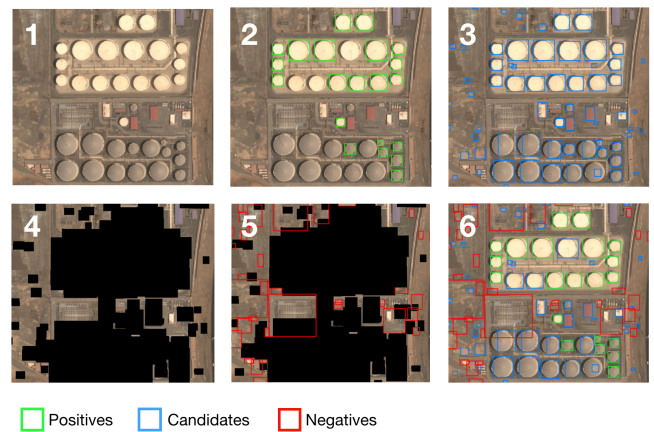


Fig. 2. Generation of positive and negative examples. The candidates are masked out prior to the extraction of the negatives. The buffer around the bounding boxes is not shown.

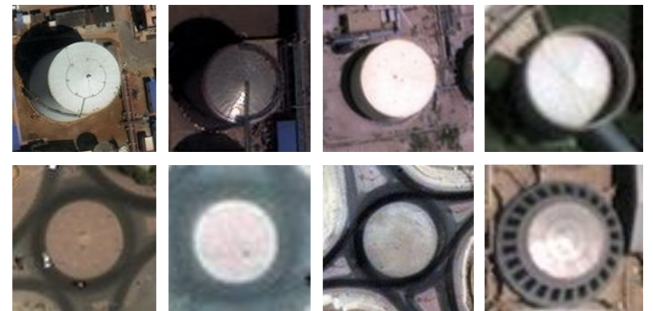


Fig. 3. A sample of automatically generated positive examples; the ones in the bottom row are false.

Asia. For each collection, we generated the orthorectified, atmospherically compensated, panchromatic and pansharpened RGB images in UTM projection. For each panchromatic image, a series of morphological operations were performed in order to isolate circular tanks from surrounding objects and to remove small patches and holes, and the max-tree and the min-tree were computed. The attribute space consists of *area*, in the range $[100, 10000]m^2$, and *compactness*, in the range $[0, 1]$, where unity is the compactness of the Euclidean disk. The generation of the training examples, illustrated in Fig.2, is now described in detail.

Positive examples should include features which we are confident are circular tanks, i.e., close-to-perfect disks. First, a segmentation was obtained by selecting the components in each root path of the max-tree within the specified area range and with $compactness \geq 0.99$; in the case of nesting, the child was retained only if it was more compact than the parent and its area was at least five times smaller than that of the parent. The procedure was repeated for the min-tree in order to obtain examples of dark circular tanks. For each segment,

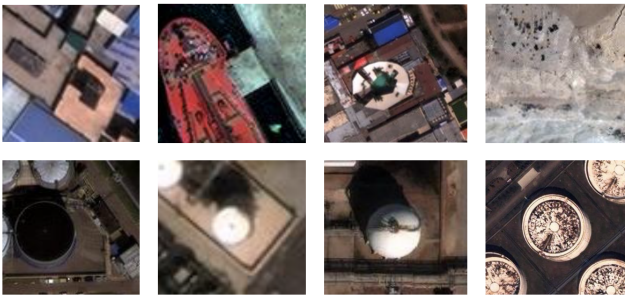


Fig. 4. A sample of automatically generated negative examples; the ones in the bottom row are false.

the axis-aligned bounding box was derived and buffered, and was used to extract the corresponding image chip from the pansharpened image; the buffering was introduced in order to provide additional context to the classifier. A few positive examples are shown in Fig.3; the bottom row consists of false positives such as roundabouts and domes. In the manner described here, we obtained 4322 positive examples, out of which we randomly selected 1250.

Following the extraction of the positives, we used the same procedure to extract the candidate bounding boxes, with the only difference that the minimum compactness value was set to 0.65. The candidates include all the positives, in addition to tanks which may have lower compactness due to the presence of rust, dirt or shadows; it may also include other objects such as buildings and patches of dirt which satisfy the minimum compactness criterion. The candidate bounding boxes are then used *to mask out the candidates from the panchromatic image prior to the extraction of the negative examples*. The reason for doing this is to avoid nesting of negative examples in tanks present in the candidates, which may occur due to the presence of non-compact components corresponding to rust, dirt or shadows within the tank perimeter. The negative examples were then obtained by computing the max-tree for the masked panchromatic image, filtering for components with compactness in the range $[0.5, 0.7]$, binarizing the output, deriving and buffering the bounding boxes of the segments, and using the latter to extract the corresponding image chips from the pansharpened image. A few samples from the negative examples are shown in Fig.4.

There is a subtlety here with regards to the compactness range used to derive the negatives. This range has to be such that the classifier is shown a sufficient number of negative examples *that look like tanks but are not*. Should a very low compactness range be used for the negatives, negatives would comprise mostly very elongated features. The classifier would then learn that a tank is a compact object, and anything else is not, and would not be able to distinguish tanks from other compact objects in the candidates.

Due to the overlap of the attribute ranges that define can-

didates and negatives, the probability that a tank ends up in the negatives, and hence contaminates the training data, increases; see the bottom row of Fig.4 for false negative examples. Since tanks are a small minority of the negatives, we can keep their number low by only training the classifier on a subset of the negatives. Out of the 586000 negative examples, we randomly selected 3750, i.e., three times the size of the positive examples, for a training set of total size 5000.

In order to assess the impact of the noise present in the automatically generated training set on the classifier accuracy, we manually corrected the false negatives and positives to obtain a clean training set. We then took the VGG-16 neural network with ImageNet initialized weights, replaced the softmax layer (ImageNet includes 1000 classes while our use case includes only two), and only trained the final convolutional block, the fully connected layers and the softmax layer until convergence, using the initial and curated training sets separately. We obtained two models, henceforth referred to as model 1 and model 2, respectively. The learning rate was set to 0.0001 and the l2 normalization factor to 0.01. The training time for each model was approximately one hour.

4. DEPLOYMENT

Two collections, one over Fujairah, United Arab Emirates, and another over Gary, Indiana were selected. A tight, axis-aligned bounding box was manually drawn around each circular tank present to create a reference data set.

We applied our detection framework in each area. This consisted of extracting from the pansharpened image the image chips corresponding to the buffered candidate bounding boxes, and feeding those to models 1 and 2 for classification. The detection set for each model consists of the unbuffered bounding boxes of the candidates that were classified as positive, where the classification was made by comparing the confidence score to a given threshold.

In order to evaluate the detector precision and recall, we calculated the intersection over union (IoU) of each detection box with each box in the reference data set and vice versa. A detection box was counted as a true positive if there was at least one reference box for which IoU was ≥ 0.5 , and a reference box was counted as detected if there was at least one detection box for which the IoU was ≥ 0.5 . Using this criterion, the precision, recall and F1 score of models 1 and 2, for the threshold values which maximized the F1 score in each case, were calculated. The results are listed in Table 1. For comparison purposes, the last row lists the corresponding accuracy metrics when the candidate set is used as the detection set (where the candidates were extracted using the compactness threshold which maximized the F1 score on the test set).

The main takeaway from Table 1 is that model 1 achieves better accuracy than compactness-based thresholding, validating the proposed framework. A visual example over Fu-

method	threshold	precision	recall	F1 score
model 1	0.999	0.734	0.728	0.731
model 2	0.800	0.921	0.722	0.809
candidates	0.990	0.935	0.521	0.669

Table 1. Accuracy metrics.

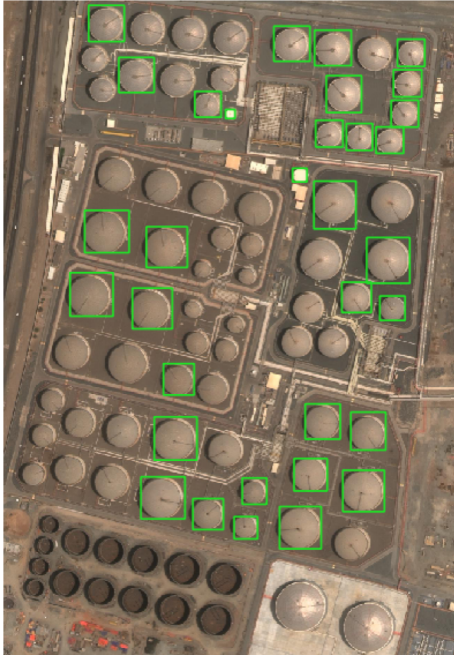


Fig. 5. Candidates for compactness threshold 0.99.

jairah is provided in Figs.5-6, where the improvement in recall, without sacrificing precision, is apparent. With regards to the impact of training data curation on the accuracy, model 2 is more precise than model 1. This is due to the fact that the positive examples in the curated data set only include circular tanks. In terms of recall, the performances are very similar, which indicates that the presence of tanks in the negative examples used to train model 1 does not really make a difference. The main reason for the relatively low recall for both models is that a lot of tanks have compactness lower than 0.65, so they are not present in the candidate set in the first place.

5. REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Electronic Proceedings of the Neural Information Processing Systems Conference*, 2012.
- [2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, Washington DC, USA, 2014, pp. 580–587.
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779–788.
- [4] A. Oliveras P. Salembier and L. Garrido, “Antiextensive connected operators for image and sequence processing,” *IEEE Transactions on Image Processing*, vol. 7, pp. 555–570, April 1998.
- [5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” <https://arxiv.org/pdf/1409.1556.pdf>, April 2015.
- [6] J. Serra, “Image analysis and mathematical morphology. volume 2: Theoretical advances,” *London: Academic Press*, 1988.
- [7] G.K. Ouzounis and M.H.F. Wilkinson, “Mask-based second-generation connectivity and attribute filters,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 990–1004, June 2007.
- [8] G.K. Ouzounis and M.H.F. Wilkinson, “Partition-induced connections and operators for pattern analysis,” *Pattern Recognition*, vol. 43, no. 10, pp. 3193–3207, 2010.

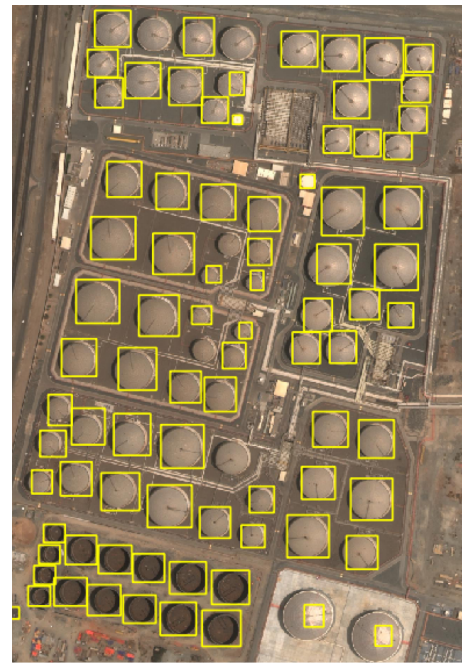


Fig. 6. Model 1 detections. Note the number of additional tanks detected compared to Fig.5.

URBAN BASELINE CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORKS ON SENTINEL-2 IMAGES

Maria Kesa⁽¹⁾, Eleni Kroupi⁽¹⁾, Victor Navarro⁽¹⁾, Camille Pelloquin⁽¹⁾, Bahaaeddin Alhaddad⁽²⁾, Laura Moreno⁽¹⁾, Aureli Soria-Frisch⁽¹⁾

(1) Starlab Barcelona SL, (2) Starlab Limited

ABSTRACT

Currently, analysing satellite images requires an unsustainable amount of manual labour. Semi-automatic solutions for land-cover classification on satellite images entail the incorporation of expert knowledge. In order to increase the scalability of the built solutions, methods that automate the image processing and analysis pipeline are required. Here we consider the task of land-cover classification of satellite images. This is an opportunity for machine learning due to the high dimensionality of the data and the need to capture the dependencies between pixels that characterize of a particular class. Recently, Deep Learning models have been applied to challenging vision problems with great success [1]. We expect the application of Deep Learning models to outperform shallow networks and other classification algorithms, as recently achieved by deep learning approaches on satellite images [2],[3],[4]. We develop a pipeline for analyzing satellite images using a deep convolutional neural network (DCNN) for use in practical applications. We present its successful application for urban classification, where it achieves 86% classification accuracy.

Index Terms— Urban classification, Segmentation, Deep Learning, Convolutional neural networks

1. INTRODUCTION

Baseline urban classification and change maps are Earth Observation products with a high added value for post analysis of urban tissue and associated urban planning activities. Currently, the semi-automatic methods providing such products require the intervention of a qualified professional for post-validation and updates. The objective of the presented work is to develop an innovative classification framework based on Deep Learning to produce the same products with, at least, the same accuracy, through a fully automatic process. The paper is presenting the products associated to the current land-cover classification, the current method and its limitations, and the new potential methods that are expected to produce more scalable systems.

2. BASELINE URBAN CLASSIFICATION MAPS

The Baseline Urban Classifications (BUC) maps provide geolocated visual data of urban land use (Fig. 1) such as arti-

cial surfaces, non-artificial surfaces and other natural and semi-natural areas. The broadest level of categorization (Level I) distinguishes among land-cover types: urban, agricultural, forest, water, irrigated lands, etc. For urban land, the second level of categorization (Level II) distinguishes among thematically detailed land uses: high and medium dense and discontinues urban fabric, main and country road and rail network.



Fig. 1 Baseline urban classification (left), Land Use change (Right), Gyumri, Armenia, from the ESA SCUDA project

BUC maps can be further processed. As an example, the Land Use Changes (LUC) maps in Fig.1 provide combined land use maps of urban classification areas over two points in time (2002 and 2014), which details the spatial characteristics of settlements evolution within the Areas of Interest.

3. CURRENT METHODS

One of the most used approaches is semantic segmentation, which is implemented through several stages. First, the image is automatically segmented into regions that fulfil a particular homogeneity criteria. Second, some segments are assigned to classes manually. Lastly, these manually labelled data is used as training set for different algorithms, i.e. K Nearest Neighbour (KNN), Support Vector Machine (SVM), or Principal Components Analysis (PCA) that classify the pixels of the entire image. The accuracies from the SCUDA project (see Fig 1) for the specific city of Vanadzor were 95.69% (producer's accuracy) and 32.84% (user's accuracy) and 0.58 kappa-score. In this context it is worth pointing out that the classification methods commonly used in image analysis practice typically present a lower performance than modern methods based on deep learning (e.g. in machine

vision competitions, such as the eminent ImageNet [5], the top performers are deep learning algorithms). We therefore expect that adopting deep learning in the image analysis pipeline will significantly improve the obtained accuracies while decreasing the manual labelling of segments in the second stage.

4. DEEP LEARNING FOR SATELLITE IMAGES

Basu and colleagues [2] use deep belief networks with feature engineering for satellite image classification. On the SAT-4 dataset the best network produces a classification accuracy of 97.95% on 500,000 image patches covering four broad land cover classes. It produced 11% better classification accuracy than state-of-the-art object recognition algorithms, Convolutional Neural Networks and Stacked Denoising Autoencoders. Given these results are large enough for ensuring its operational application, our aim is to increase the classification accuracy also for other types of satellite images, e.g., in the Sentinel-2 dataset. We will minimize in this manner the human labour involvement in the classification chain.

Another trend in the land-cover classification procedure for satellite images are the semi-supervised learning approaches that have been applied with excellent results [6]. The methodology allowed to reduce the labelling time of data from approximately 9 hours to 30 minutes. The semi-supervised approach achieved 87.93% compared to 91.38% (Sydney dataset) and 86.83% compared to 92.50% (Washington dataset) fully supervised model accuracy. Their method used Discriminative Sparse Autoencoders for extracting high-level features from data. While this model performed really well, we believe that there are benefits in exploring the deep learning model space for semi-supervised learning further, so that different algorithms could be benchmarked against each other. The literature in applying deep learning algorithms in a semi-supervised fashion on satellite images is sparse, perhaps because deep networks require a huge amount of training data, which is usually not the case in semi-supervised approaches. There is some older work which employed neural networks (see [7], [8]), but modern deep learning approaches are only beginning to be explored by the community. Thus, we target moving towards semi-supervised classification using DCNNs, benefiting from the advantages of both the deep learning frameworks and the semi-supervised ones. However, and in order to achieve this final goal, we need first to evaluate the performance of the DCNNs trained on an already existing database of satellite images that contains ground truth, and test it on an unseen image of the same satellite type. Having a model already trained and being able to directly recall it to classify new images of the same satellite type will already reduce the time spent in land-cover classification in a semi-automated way. This is the work we present in this paper. The image used as the test set was previously classified in a semi-automated manner using

the semantic segmentation approach, which is described in the former section. Our final goal is to achieve good classification performance in this image, already by training the DCNN's with an existing database and applying the model to the new image. We plan to eventually improve the performance by including in the algorithm patches from the target image in a semi-supervised way.

5. METHODS

Our images and the problem we aim to solve are derived from the Sentinel-2 high-resolution multispectral satellite imager, which presents 13 spectral bands. However in this study we use four bands, namely RGB and near infrared to simplify the problem. The procedure followed is summarized in the following steps. Initially, the AlexNet model that has been shown to provide very high classification accuracies in the DeepSat database was implemented in tensorflow following [9]. Briefly, the model contains seven types of layers, namely convolutional, RELU, maxpool, drop-out, a threshold layer, fully connected layers, and a final softmax layer. The layers are arranged sequentially. The only change made with respect to the initial model is the use of the Adam optimizer that already includes learning rate decay, so the exponential learning rate decay used in [9] has not been implemented. Once the model was implemented, it was trained and tested in the DeepSat database, for validation purposes. The results are presented in Section 6.

The DeepSat database contains images of very high resolution (1-6m per pixel) compared to the Sentinel-2 images (10m per pixel). Thus, the first challenge was to explore whether the usage of the DCNN model can be extended to successfully analyse (i.e. achieve accuracy higher than 80%) also other types of satellite images (e.g., Sentinel-2 images). Hence, the DCNN model was also trained and tested in the newly published EuroSat database [10], which consists of a subset of Sentinel-2 satellite images with their ground truth (GT). The GT covers ten classes, namely Industrial, Residential, Annual crop, Permanent crop, River, Sea & Lake, Herbaceous vegetation, Highway, Pasture, and Forest. We performed object-based classification, i.e. we segmented the images into patches, as in [9-10]. Although the patch size in [10] was of 64x64 pixels, we used 16x16 size patches to get the result into a finer resolution.

Once a good classification performance was achieved (see Section 6), we trained the model using all data from the EuroSat database and tested it in a Sentinel-2 image of Yangon city and surroundings in Myanmar that we had previously classified at pixel level, in a semi-automatic way through semantic segmentation (see Section 3). The rationale was to investigate how the model performed in an unseen image of the same satellite. Our ground-truth contained nine classes, namely Urban fabric, Main and country road, High dense green area – Forest/Trees areas,

Medium dense green area – Trees/Agricultural Areas, Low dense green area - Open/Urban green areas, Irrigated field areas, Semi-irrigated field areas, Open space with little or no vegetation, and Shallow water. As a first step, we grouped manually the classes of both the EuroSat database and the Yangon city image and map them into three final classes, namely urban, non-urban and water. Our classification of the Yangon city served as the GT for this image. Since the GT of the Yangon city was pixel-wise, we created 16x16 patches and used majority voting on the pixels of each patch to infer the patch GT.

For both datasets, the images were not environment-corrected to explore the performance using the raw data, which in operational conditions would save time from the processing chain. However, all training images were standardized subtracting the mean and dividing by the standard deviation pixel-wise. The same procedure was applied also in the test data, using the mean and standard deviation obtained from the training data. For all datasets 4 bands were used, namely RGB and near-infrared.

In order to visualize the outcome, we output the final probabilities, instead of the crisp class labels. Here we map the probability of each of the three classes into one colour channel value for each patch. This type of representation allows to visualize the possible uncertainties in the final classification map. If, for instance, a patch presents a probability membership vector $[0.5, 0.4, 0.1]$, where e.g. 0.5 is the probability of the patch belonging to the first class, that patch is represented through RGB colour values $[127, 102, 25]$. This RGB tuple would better represent the associated decision uncertainty than the tuple $[255, 0, 0]$ that corresponds to the crisp presentation.

6. RESULTS

Regarding the DeepSat database, we trained the DCNN model on 234000 images of 28x28 patches and tested it on 70000 images. The final classification accuracy was 93.2%, verifying that the model has been well implemented.

The next step was to apply the model in the newly developed EuroSat dataset (Sentinel-2 images dataset). In this case we used 354600 images of 64x64 patches for training and 10000 images for testing. In order to minimize a wrong performance estimation due to the training set variability, we randomly selected the training and test sets and repeated this procedure three times. The average classification accuracy and its standard deviation are presented in Table 1, for each class and as a total.

Table 1. Classification results from the EuroSat database with DCNN

Class Name	Average accuracy (%)	Standard deviation
Annual crop	91.3	2.2
Forest	98.6	0.8
Herbaceous vegetation	91.8	2.1

Highway	61.5	3.1
Industrial	84.4	5.5
Pasture	90.8	1.3
Permanent crop	85.8	4.9
Residential	93.2	1.2
River	80.8	1.5
Sea Lake	97.4	0.3
Total average accuracy	87.6	2.3

One may notice from Table 1 that the average classification accuracy is high ($>80\%$), which implies that the DCNN framework implemented for the DeepSat database with high resolution can be also extended to the EuroSat database (Sentinel-2 data) that has lower resolution. The classification accuracy drops, but the result is still very high. Another observation made from Table 1 is that the Highway class is often misclassified, and mainly confused with Herbaceous vegetation and Industrial, as observed from the confusion matrix presented in Fig 2.

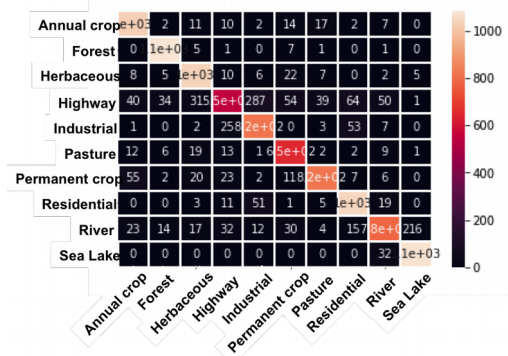


Fig. 2 Confusion matrix for the EuroSat dataset

As previously described, the final performed test was to train the model using the EuroSat images and test it using an image from the Yangon city that we had previously classified and was not included in the training set. The image is presented in Fig 3a and the GT in nine classes is presented in Fig 3b.

The final classification accuracy results for this case are presented in Table 2. The *accuracy* refers to the correctly classified number of patches with respect to the total number of patches, whereas the *mean accuracy* refers to the average accuracy across the three classes. The accuracy is lower than the mean accuracy since the class Non-urban contains many more patches compared to the rest of the classes, and is the one with the lowest performance.

Table 2. Classification results training the DCNN on EuroSat and testing it on Yangon city image

Class name	Accuracy (%)
Non-urban	63.6
Urban	95.8
Water	94.6
Accuracy	71.2
Mean accuracy across classes	84.6

The confusion matrix for this case is presented in Fig 4. It is noticeable that the Non-urban class is misclassified as urban, but as water as well. The classes Water and Urban are very well classified (more than 90% accuracy). The final image with the classes identified based on their probabilities is presented in Fig 3c.

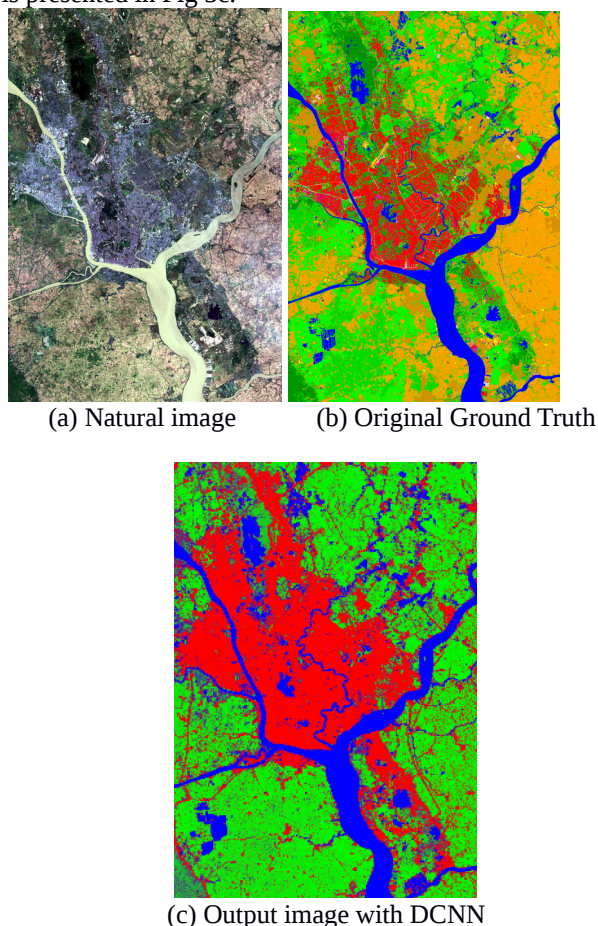


Fig. 3 (a) The natural image from the Yangon city and its surrounding (Sentinel-2), (b) the estimated GT (9 classes), and (c) the classification result in 3 classes using the DCNN.



Fig. 4 Confusion matrix for the Yangon city

Visual inspection of the Fig 3b and 3c allow us to verify that the land cover has been successfully carried out using DCNN, although there is still room for improvement. One may note that the initial GT was developed in a semi-

automatic way, which implies that is not 100% accurate. This may mean that maybe some points have been better classified with DCNN compared to the provided GT, which would require further investigation.

7. CONCLUSIONS

In this paper we have reviewed how the work on satellite images is currently carried out and outlined a research problem that aims to improve the current state of affairs. We have implemented a DCNN model that was previously tested on the DeepSat database [9]. This DCNN has been successfully validated and tested on the EuroSat dataset that consists of Sentinel-2 images. We achieved a mean accuracy of 87.6% in this problem. We then trained the same model using the whole EuroSat dataset and tested it in a Sentinel-2 image, which we had previously classified in a semi-automatic way. The average classification result for a three-class problem was 84.6%, indicating that the DCNN model constitutes a promising tool for automatic land-cover classification of Sentinel-2 images. In our future work we plan to include some patches from the target image in the training of the model. Finally we expect from the semi-supervised extension of the methods an improvement in the overall classification accuracy and therefore its successful translation into operational applications.

8. REFERENCES

- [1] C. Szegedi, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going Deeper With Convolutions", IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.
- [2] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, R. Nemani, "DeepSat-- A Learning Framework for Satellite Imagery", ArXiv, 2015.
- [3] M. Langkvist, A. Kiselev, M. Alirezaie, A. Loufi, "Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks", *Remote Sensing*, 8(4), 329, 2016.
- [4] V., Mnih, "Machine Learning for Aerial Image Labeling", PhD thesis, University of Toronto, 2013.
- [5] ImageNet competition, <http://image-net.org/>
- [6] X. Yao, J. Han, G. Cheng, X. Qian, L. Guo, "Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning", IEEE Transactions on Geoscience and Remote Sensing, 2016.
- [7] G. Camps-Valls, D. Tuia, L. Bruzzone, J.A. Benediktsson, "Advances in Hyperspectral Image Classification- Earth Monitoring with Statistical Learning Methods", ArXiv, 2013.
- [8] F. Ratle, G. Camps-Valls, "Semi-supervised Neural Networks for Efficient Hyperspectral Image Classification", IEEE Transactions on Geoscience and Remote Sensing, 2009.
- [9] M. Papadomanolaki, M. Vakalopoulou, S. Zagoruyko, K. Karantzas. 2016. "Benchmarking Deep Learning Frameworks for Classification of Very High Resolution Satellite Multispectral Data", ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume III-7.
- [10] P. Helber, B. Bischke, A. Dengel, D. Borth. 2017. "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification", ArXiv

SPUSPO: SPATIALLY PARTITIONED UNSUPERVISED SEGMENTATION PARAMETER OPTIMIZATION FOR EFFICIENTLY SEGMENTING LARGE HETEROGENEOUS AREAS

S. Georganos¹, T. Grippa¹, M. Lennert¹, S. Vanhuysse¹, E. Wolff¹

¹ Université libre de Bruxelles (ULB), Department of Geosciences, Environment and Society (DGES-IGEAT)

ABSTRACT

Very-High-Resolution (VHR) Remote Sensing (RS) data are crucial for deriving essential geospatial information on cities, e.g. for urban planning, population estimation and socioeconomic assessments with particular merit in sub-Saharan Africa (SSA) due to the scarcity or absence of reference data. One of the cornerstones of information that can be produced from RS is classified Land Use and Land Cover (LULC) maps. For VHR imagery, Object Based Image Analysis (OBIA) is the most efficient methodology to produce such outputs. A crucial intermediate step in OBIA is the selection of a suitable segmentation scale. However, for large, heterogeneous areas (e.g., at city level), little effort has been made to optimize OBIA algorithms. Supervised methods to optimize segmentation parameters are subjective and time consuming while Unsupervised Segmentation Parameter Optimization (USPO) techniques, assume spatial stationarity for the whole image. This is problematic for geographically large heterogeneous areas and does not capture intra-urban variations due to building size, materials and fractions of LULC intrinsically varying in space. In this study, we employ a novel framework named Spatially Partitioned Unsupervised Segmentation Parameter Optimization (SPUSPO) that optimizes segmentation parameters locally for two SSA cities, Dakar and Ouagadougou. The framework employs the open access GRASS GIS software that is suitable for large scale computing. Our results suggest that SPUSPO is an efficient way to optimize segmentation parameters for large and heterogeneous urban areas.

Index Terms—unsupervised segmentation parameter optimization, GRASS GIS, VHR imagery, land cover

1. INTRODUCTION

In sub-Saharan African cities, in the absence of adequate data for urban planning, one of the most important pieces of information we can derive through Very-High-Resolution (VHR) Remote Sensing (RS), is detailed and accurate Land Use and Land Cover (LULC) maps. These products are often used as input for epidemiological, population and socio-economic models, among others. Nonetheless, to achieve adequate classification results in these challenging landscapes, Object Based Image Analysis (OBIA) [1] is frequently employed over pixel-based approaches due to superior performance [2]. Nonetheless, OBIA can be a very tedious and computationally demanding technique, particularly for large-scale applications (e.g., at city level).

Moreover, the classification of relevant images requires several preparatory tasks, ranging from the selection of an appropriate segmentation (object-creating) algorithm, to the effective calibration of the parameters of the algorithm itself [3]. Such parameters control the shape and size of the created segments. Since the segmentation quality is crucial to the results of the classification itself, the optimization of these parameters is a critical methodological facet.

Region-growing segmentation algorithms are very frequent in the OBIA literature, mainly due to their effectiveness and ease of implementation. The selection of the parameters of the segmentation algorithm is most commonly achieved through a time-consuming, user dependent, trial and error process in which the quality of the segmentations is assessed visually or through alternative methodologies which rank different segmentations based on reference data. However, supervised calibration approaches are untenable for large heterogeneous regions if maximizing classification accuracy and segmentation quality in an automated approach is one of the aims. Consequently, research efforts have been directed towards the development of objectively defined Unsupervised Segmentation Parameter Optimization (USPO) techniques, that evaluate individual segmentations based on geostatistical metrics, and do not require reference data and user interference, thus allowing for automated processes [4]. The optimization of USPO metrics has been attempted mainly by the use of global methods, either at single or at multiple scales. This means that the segmentation over the whole region of interest is optimized by using one set of parameters. This global approach has recently been shown to be inferior to local approaches in urban and agricultural environments either through evaluation metrics such as classification accuracy and detailed visual examination [5],[6]. Thus, in order to treat the growing amount of VHR remote sensing data available in a way that is both time-efficient and accurate, an approach based on the hypothesis that the segmentation optimal parameters would intrinsically and significantly vary across heterogeneous regions due to local variations in data structure, is more appropriate. However, there is no established automated framework for applying local USPO for very large VHR urban datasets. In general, little effort has been made to identify and quantify the degree of spatial non-stationarity between the algorithm parameters, especially for large heterogeneous areas such as cities, where neighborhood structure and consequently, optimal segmentation parameters might vary due to different building materials, size, fractions and types of vegetation

and road network. As such, our aim is to present a methodological framework to optimize segmentation results for large heterogeneous areas based on the following premises: i) allowing segmentation parameters to vary, which can provide significant increases in classification quality and ii) implement the methods through an open source semi-automated processing chain, suitable for large-scale computing utilizing mainly GRASS GIS, in order to make the previous task manageable.

2. DATA AND STUDY AREA

The classification scheme is implemented for Dakar and Ouagadougou, the capitals of Senegal and Burkina Faso, respectively (Figure 1). They are both major Sahelian cities, which have been facing an extensive urban growth since the last decades. Pansharpned tristereo Pleiades imagery (VNIR, 0.5m, 400 km²) acquired in 2015 was used for Dakar while Worldview-3 stereo imagery acquired in the same year was used for Ouagadougou (VNIR, 0.5m, 630 km²). Normalized Digital Surface Models (nDSM) were produced by photogrammetry from both datasets to aid in built-up fabric identification.

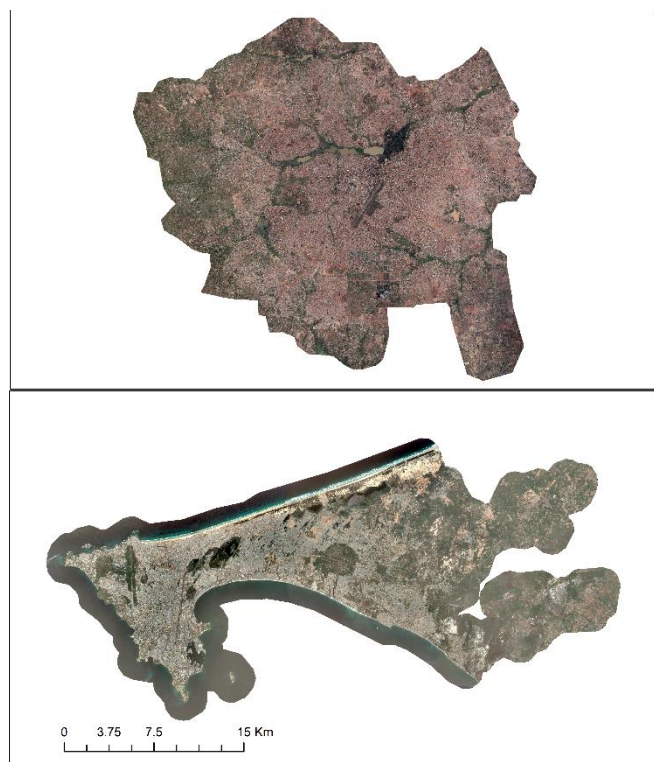


Figure 1. True color composites of Ouagadougou (top) and Dakar (bottom).

3. METHODS

In this paper, we present a methodological framework named Spatially Partitioned Unsupervised Segmentation Parameter Optimization (SPUSPO) in which

optimization of segmentation parameters is performed locally. We followed a different approach for the spatial partitioning of each city, a fully unsupervised process for Dakar and a supervised one for Ouagadougou (Figure 2). For Dakar, we split the image into 1700 500m by 500m grid cells. For Ouagadougou, an expert-based visual delineation of local morphological zones (LMZ) was performed by applying the following ruleset:

- LMZs should be homogeneous, both in terms of building size and density, and should be visibly different from their neighboring LMZs.
- LMZs boundaries should follow, as far as possible, man-made or natural linear elements, e. g., roads, paths, rivers, streams, railways.
- Built-up LMZs should be larger than 1.5 hectares (ha).

Both of these spatially partitioning methods were compared to a global approach in which segmentation parameters were optimized in a small subset of the study areas and the suggested value was applied to the whole study area.

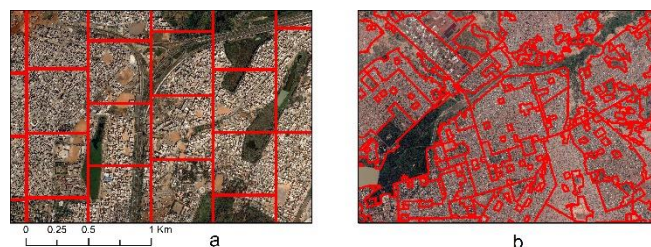


Figure 2. Different methods of spatial partitioning **a)** automatic delineation of grid cells in Dakar and **b)** manual, expert delineation of LMZ in Ouagadougou.

Afterwards, we made extensive use of the semi-automated processing chain developed by Grippa et al. [7], that combines GRASS GIS [8], Python and R programming languages along with PostGIS support in a Jupyter Notebook implementation. The chain allows for a complete analysis, from input of the initial image datasets to the production of final LC maps. An excerpt example of the Jupyter chain is given in Figure 3, while the general workflow is described in Figure 4. In detail, the segmentation was optimized in each of the spatial subsets (grid cells for Dakar and LMZs for Ouagadougou) using the *i.segment.uspo* module of GRASS [9] with a region growing algorithm. The module uses geospatial metrics to perform the optimization, such as Moran's I and the variance to find the best compromise between intra- and inter-segment heterogeneity. Moreover, it allows for parallelization and is thus computationally efficient. Once layers of features calculated on the segmentation layers coming from different techniques were produced, a random forest classifier was implemented to evaluate their efficacy [10]. To train and evaluate the models we visually identified roughly 3000 objects for both cities. As shown in Table 1, we used a detailed classification scheme to draw more robust conclusions about the merits of using local optimization on

large regions.

```

Set list of raster from which to compute statistics with i.segment.stats
Please refer to the official help page of i.segment.stats to select the raster statistics and area measures to be computed.

In [271]: # Display the name of rasters available in PERMANENT and CLASSIFICATION mapset
print grass.read_command('g.list',type='raster', mapset='PERMANENT', flags='rp')
print grass.read_command('g.list',type='raster', mapset-user['classificationC_mapsetname'], flags='rp')

raster fichiers disponibles dans le jeu de données <PERMANENT> :
NEXT      NEXT      531      ndsm      opt_blue  opt_green opt_nir  opt_red

-----
raster fichiers disponibles dans le jeu de données <CLASSIF> :
segments  zone_morpho

Compute statistics of segment using i.segment.stats
The process is make to compute statistics iteratively for each morphological zones, used here as tiles.
This section uses the i.segment.stats_asis-on to compute statistics for each object.

In [272]: # Save name of the layer to be used as tiles
tile_layer='zone_morpho'@mapsetname
# Save name of the segmentation layer to be used by i.segment.stats
segment_layer='segment'@mapsetname
# Save name of the column containing area_4m value
area_column='area_4m'
# Save name of the column containing morphological type value
type_column='type'
# Save the prefix to be used for the outputfiles of i.segment.stats
prefix='segment_stats'

In [281]: # Save the list of polygons to be processed (save the 'cat' value)
listofregion=list(grass.parse_command('v.db.select', map=tile_layer,
                                     columns='cat', flags='c'))

In [279]: count=1
for cat in listofregion:
    condition='cat=%cat'
    typemorph=grass.read_command('v.db.select', map=tile_layer,
                                columns='type_column, where=condition, flags='c')

```

Figure 3. Excerpt of the Jupyter notebook consisting of a sequence of code and a descriptive text that documents the different processing steps that can be executed directly from the notebook.

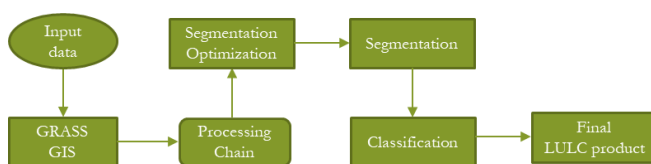


Figure 4. General workflow of the processing chain.

Table 1. Classification scheme for Ouagadougou and Dakar

Ouagadougou	Dakar
Buildings	Building type 1
Swimming Pools	Building type 2
Asphalt	Asphalt
Brown/Red Bare Soil	Bare soil, dusty concrete
White/Grey Bare Soil	
Trees	Trees
Mixed Bare Soil/Vegetation	Low vegetation (e.g. grass)
Dry Vegetation	Bushes
Other Vegetation	
Inland Waters	Inland waters
Shadow	Shadow

4. RESULTS

The results indicated that a local approach can be of merit.

Figure 5 demonstrates the variability of optimized *threshold* parameter for each grid cell across Dakar (Figure 5). It is evident that the local method suggests spatial non-stationarity of the USPO parameter values. The patterns of the deviation from the single parameter of the global approach follow variations in the landscape such as the types of vegetation and built-up fabric. Supporting our hypothesis, the results from evaluating the classification against a reference set implied that undertaking a local optimization path constructs segments of a higher quality. The overall accuracy (OA) for Dakar reached 88.20% and 89.50% for the global and local approaches, respectively. In Ouagadougou, similar results were observed with an OA of 84.77% and 85.45%. Being an urban application, it is also relevant to examine how well the built-up classes are predicted. By using the F-score as an indicator to evaluate per class performance, the results indicate an increase of 2% and 1% for Dakar and Ouagadougou, respectively. Although the improvements are not very strong, a detailed visual examination implied that the local approaches predict built-up and asphalt classes, in a more consistent fashion. This can be important given that these products are frequently used for population and epidemiological applications where built-up information is the most important feature.

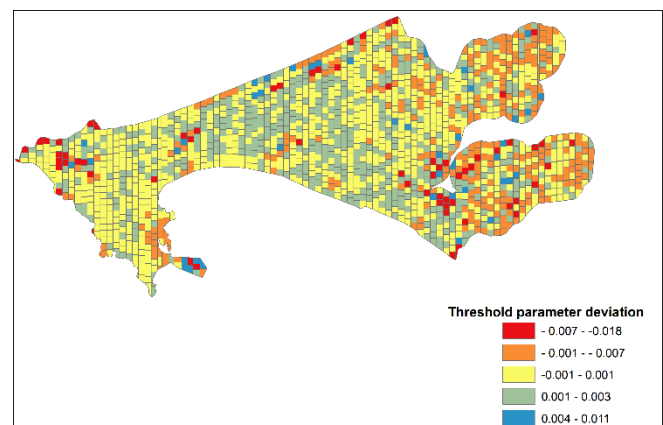


Figure 5. Deviation of the local threshold parameter from the global one for each grid cell in Dakar. The threshold parameter controls the shape and size of the segments and thus, the quality of the segmentation.

There are several reasons that could explain these improvements. A single segmentation parameter can over-segment and under-segment objects of the same class in different regions (Figure 6). For example, industrial areas consist of large buildings while unplanned areas contain very small housing units. By allowing the threshold parameter to vary, SPUSPO can fit these changes in the data allowing for better classification results.

5. CONCLUSIONS

In this paper, we assessed a new method for optimizing segmentation parameters in large and highly heterogeneous

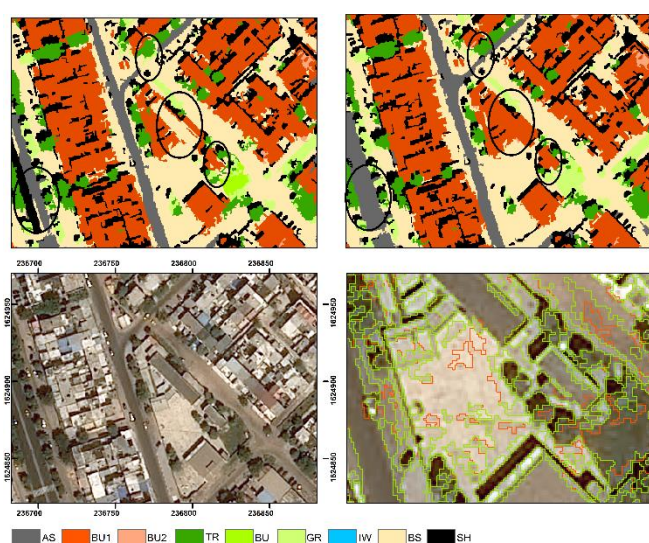


Figure 6. Segmentation and classification results for Dakar. Top left: Classified segments resulting from the global approach. Top right: Classified segments resulting from the local approach. Bottom left: True color composite. Bottom right: segments boundaries coming from the global (red) and local (green) approaches.

Table 2. Classification accuracy for Ouagadougou and Dakar with a global approach and SPUSPO, respectively.

	Local		Global	
	Ouagadougou	Dakar	Ouagadougou	Dakar
OA%	85.45	89.50	84.77	88.20
Kappa%	0.840	0.878	0.833	0.863
Built-up%	0.930	0.960	0.920	0.940

urban areas with the prospect of deriving high quality LU/LC maps. We call this method Spatially Partitioned Unsupervised Segmentation Parameter Optimization (SPUSPO). We employed two different ways to partition the images i.e., automatic partition by grid cells and expert based delineation using morphological criteria. The methodological framework was realized through an open source processing chain that mainly exploited the artillery of GRASS GIS for large scale computing. Our results suggested that spatial partition methods have merit as increased classification accuracies were observed in comparison to a global approach, along with better prediction of artificial surfaces. Future work includes more testing of the method in different imageries and refining techniques of automatic delineating useful spatial subsets to optimize the data. Finally, a comparison between the two approaches demonstrated in this study— automated vs expert based, will be attempted to assess if a completely automated approach can be as efficient with a focus in built-up areas

identification as they are the most challenging elements to classify in SSA cities.

ACKNOWLEDGEMENTS

The research presented in this paper is funded by BELSPO (Belgian Science Policy Office) in the frame of the STEREO III program – project REACT (SR/00/337).

6. REFERENCES

- [1] T. Blaschke, “Object based image analysis for remote sensing,” *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 2–16, 2010.
- [2] T. Blaschke *et al.*, “Geographic Object-Based Image Analysis - Towards a new paradigm,” *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 180–191, 2014.
- [3] A. Rasanen, A. Rusanen, M. Kuitunen, and A. Lensu, “What makes segmentation good? A case study in boreal forest habitat mapping,” *Int. J. Remote Sens.*, vol. 34, no. 23, pp. 8603–8627, 2013.
- [4] B. Johnson, M. Bragais, I. Endo, D. Magcale-Macandog, and P. Macandog, “Image Segmentation Parameter Optimization Considering Within- and Between-Segment Heterogeneity at Multiple Scale Levels: Test Case for Mapping Residential Areas Using Landsat Imagery,” *ISPRS Int. J. Geo-Information*, vol. 4, no. 4, pp. 2292–2305, 2015.
- [5] T. Grippa, S. Georganos, M. Lennert, S. Vanhuyse, and E. Wolff, “A local segmentation parameter optimization approach for mapping heterogeneous urban environments using VHR imagery,” in *Remote Sensing Technologies and Applications in Urban Environments II*, 2017, vol. 10431, p. 104310G.
- [6] F. Cánovas-García and F. Alonso-Sarría, “A local approach to optimize the scale parameter in multiresolution segmentation for multispectral imagery,” *Geocarto Int.*, vol. 30, no. 8, pp. 937–961, 2015.
- [7] T. Grippa, M. Lennert, B. Beaumont, S. Vanhuyse, N. Stephenne, and E. Wolff, “An Open-Source Semi-Automated Processing Chain for Urban Object-Based Classification,” *Remote Sens.*, vol. 9, no. 4, p. 358, 2017.
- [8] M. Neteler, M. H. Bowman, M. Landa, and M. Metz, “GRASS GIS: A multi-purpose open source GIS,” *Environ. Model. Softw.*, vol. 31, pp. 124–130, 2012.
- [9] M. . Lennert and G. D. Team, “Addon i.segment.uspo,” *Geogr. Resour. Anal. Support Syst. SoftwareVersion 7.3*, 2016.
- [10] S. Georganos, T. Grippa, S. G. Vanhuyse, M. Lennert, and E. Wolff, “Optimizing classification performance in an object-based very-high-resolution land use-land cover urban application,” in *Remote Sensing Technologies and Applications in Urban Environments II*, vol. 10431.

A GLOBAL MOSAIC FROM COPERNICUS SENTINEL-1 DATA

V. Syrris¹, C. Corbane², and P. Soille¹

European Commission, Joint Research Centre (JRC)

¹Directorate I. Competences, Unit I.3 Text and Data Mining,

²Directorate for Space, Security & Migration, Unit E.1 Disaster Risk Management,
via E. Fermi 2749, I-21027 Ispra (VA), Italy

ABSTRACT

This paper presents an algorithmic workflow for producing mosaics based on the dual polarisation capability of Sentinel-1 SAR imagery. The main characteristics of the specific method are: automated and nonparametric approach, fast processing, incremental adjustment and information distinction. The workflow has been optimized according to the configuration of the recently introduced JEODPP platform. Challenges, suggestions and solutions are discussed as well.

Index Terms— Mosaic, Sentinel-1, Copernicus, histogram discretization

1. INTRODUCTION

Sentinel-1 (S1) space mission is a constituent project designed and managed by the European Space Agency (ESA) within the framework of the Copernicus programme [2]. Technically, it is a constellation of two satellites, Sentinel-1A and Sentinel-1B, sharing the same orbital plane with a 180° orbital phasing difference, aiming at acquiring systematically and providing routinely data and information products to Copernicus Ocean, Land, and Emergency as well as to national user services.

Generating and assembling composites from satellite imagery is of great importance, giving birth to applications ranging from the construction of a base-layer to more sophisticated processes like the spatio-temporal signal comparison and analysis. Mosaicking is the process of stitching together image tiles which have a unique spatio-temporal stamp for generating a seamless, homogeneous canvas. Since the launch of S1A mission, various S1-mosaics have been generated at national and regional levels such as [5, 1, 3, 4].

This paper describes an algorithmic workflow for building mosaics based on S1 data, following a fully automated and nonparametric approach, suitable for being executed in a high-throughput computing facility. It can be considered equally as a demonstration of a real application supporting efficient handling and processing of big Earth observation data.

2. THE JEODPP PLATFORM

The experiments and the processing have been done on a high-throughput computational platform called the Joint

Research Centre Earth Observation Data and Processing Platform (JEODPP) [12, 13]. The main components of the JEODPP infrastructure are presented briefly below:

- storage system: based on Just a Bunch of Disks (JBODs) managed by *EOS* open-source distributed file system [9] developed and maintained by the European Organization for Nuclear Research (CERN);
- processing servers and related services: a high performance commodity cluster at which a flexible scheduler (HTCondor [8]) undertakes the task of distributing the load over processing servers connected to the storage servers administered by *EOS*;
- network system: storage and metadata servers are connected via (single or double) bonded network configuration;
- virtualisation: Docker containers [11] allow for flexible management of hardware resources and processing environments. They function as a light-weight type of virtualisation to separate processing instances.
- user interface: two web-based modes are provided to the user for fast prototyping and analytics via i) remote desktop supported by the Apache Guacamole gateway, and ii) interactive visualization and on the fly processing through Jupyter notebooks and in-house hard-coded libraries.

3. DATA

The application for which the data were designated is the Global Human Settlement Layer (GHSL) project [7] and more specifically, the generation of a global built-up map based on S1 data [10]. The dataset consists of 5,026 Sentinel-1A products (~8TB) covering almost completely the globe, spanning over the year of 2016, acquired in IW (Interferometric Wide swath) mode, processed in GRDH format, and coming in dual polarisation (mostly VV+VH). The data were queried and downloaded using OpenSearch and OData scripting capabilities offered by the Copernicus Service hub [6] operated by the ESA.

The objective of the presented work was the use of the aforementioned dataset for the generation of a homogeneous base-layer; in that way, the user can easily contrast the input S1 imagery and the GHSL output and assess better the results.

4. PROCESSING WORKFLOW

Level-1 Ground Range Detected (GRD) products are projected to ground range using an Earth ellipsoid model such as WGS84 (slant range coordinates), with pixel values representing detected magnitude. Calibration and geo-location takes place as a pre-processing step in our workflow, executed usually once per application. The main processing of image blending and mosaicking can have several flavours (different color compositions and mosaicking rules), therefore the entire processing workflow was split in two phases.

4.1. Pre-processing

The pre-processing was done via the S1 toolbox (S1TBX version 2.0.2) and comprises the following modules:

1. Apply orbit file: this file provides accurate satellite position and velocity information, based on which the orbit state vectors in the abstract metadata of the Sentinel-1 product are updated;
2. Thermal noise removal: using the noise vectors it removes dark strips near scene edges with invalid data. On the output of this process, we applied a mathematical morphology operation in order to cut off completely the uncertain borders;
3. Radiometric calibration: it computes the backscatter intensity (sigma nought) using sensor calibration parameters in the GRDH metadata;
4. Terrain correction (orthorectification): it converts data from ground range geometry into a map coordinate system. For the majority of the products we employed the SRTM 1 arc sec HGT DEM and for the remaining 548 products the ASTER 1 arc sec GDEM due to the non-exhaustive coverage of the SRTM. The resampling was done via the bilinear interpolation method.

Due to the need for bringing to light small targets like sparse settlements, we decided to omit the speckle filtering stage. Nevertheless, depending on the application, one can easily add this process (usually before step 4) since S1TBX already supports single product and multi-temporal speckle filtering.

4.2. Main-processing

The pre-processing outcome for each product is two geo-referenced images (backscatter coefficient for the two polarisations as floating-point) in WGS 84/Pseudo-Mercator coordinate system (EPSG:3857) with 19.11 spatial resolution, corresponding to the 13 zoom level in Tile Map Service (TMS). The main processing steps are as follows:

4.2.1. False color composition rules

The ability to visually differentiate built-up areas from other natural or man-made features depends on the optimal combination of bands that provides the maximum separability between those different features. The false colour composite proposed here uses the 8-bit discretization of the dual polar-

ization backscatter values as follows: the VH or HV as Red, a linear mapping over the average of (VH,VV) or (HV,HH) as Green, and the VV or HH respectively as Blue band.

1. Saturate the extreme values: Given that image $I_{X_i, i=1,2}$ denotes one of the two $\{VH, VV\}$ or $\{HV, HH\}$ polarizations with V: Vertical, H: Horizontal, apply the function

$$I_{X_i}(v) = \begin{cases} 1, & v > 1 \\ 10^{-4}, & -1 < v \leq 10^{-4} \\ v, & \text{otherwise} \end{cases}$$

where -1 has been set as no data value. Even if the range of values after calibration/terrain correction is expected to lay out between 0 and 1, few negative values may be produced due to wrong operation of the thermal noise removal which is intended for scenes over land and not over the ocean. In the other side of the range, few values greater than 1 may be attributed to local strong scatterers. Without affecting the outcome, the function above set all the values to the expected range;

2. Compute the common data domain for the two polarisations by applying morphological operators to clean further the noisy/low value borders:

$$D = \bigcap_{X_i, i=1,2} \varepsilon_{N_{7 \times 7}} \left(\delta_{N_{5 \times 5}} (D_{X_i}) \right), \text{ where } N_{7 \times 7}$$

and $N_{5 \times 5}$ are two structuring elements selected for making the area compact and for cropping the image borders sufficiently; the functions δ and ε correspond to the morphological operations of *dilation* and *erosion*. The binary images D_{X_i} have value 1 when for the corresponding values of $I_{X_i}(v)$ it holds $v > -1$ and 0 otherwise. Subsequently, the no data values of every I_{X_i} are being updated according to D ;

3. Convert each I_{X_i} to its logarithmic counterpart and discretize its values in 8-bits by binning appropriately the image values distribution: Even though the standard practice is to transform to db via the function $10 \cdot \log_{10}(\cdot)$, in this case any kind of logarithmic function is producing equivalent result. The natural logarithm spreads intuitively the values without the multiplication by an extra factor; hence, $I_{X_i}^l = \ln(I_{X_i})$. By analysing the statistical distributions (SD) of the continuous values of all the $I_{X_i}^l$, we estimated the following critical ranges and values: for cross polarizations, i) $h = [8, 124, 107, 14, 2]$, $r = \ln\left([10^{-4}, 10^{-2}, 0.035, 0.06, 0.12, 1]\right)$ when SD is close to normal, ii) $h = [8, 144, 87, 14, 2]$, $r = \ln\left([10^{-4}, 10^{-2}, 0.025, 0.06, 0.12, 1]\right)$ when SD is left-skewed, and for co-polarizations, iii) $h = [14, 122, 105, 12, 2]$, $r = \ln\left([10^{-4}, 0.04, 0.14, 0.32, 0.63, 1]\right)$ when SD is close to normal, iv) $h = [14, 142, 85, 12, 2]$, $r = \ln\left([10^{-4}, 0.04, 0.12, 0.32, 0.63, 1]\right)$ when SD is left-skewed. For each of the four cases, the respective h and r vectors steer

the recursive construction of a vector C which contains $|C| = 255$ values: $C = \bigcup_{k=1, \dots, 5} \left[r_{1,k}, r_{j+1,k} = r_{j,k} + \frac{r_{1,k+1} - r_{1,k}}{h_k - 1} \mid j = 1, \dots, h_k - 1 \right]$. Finally, by utilizing the C vector as being reversely ordered by the maximum to the minimum value, the data binning is being carried out as follows: $\forall v$ of $I_{X_i}^l : v < c_d \in C \Rightarrow B_{X_i}(v) = d$, where $d \in \mathbb{Z}^+ : 1 \leq d \leq 255$; c_d are the values of C used as thresholds over $I_{X_i}^l(v)$. $B_{X_i}(v)$ denotes the discrete representation of $I_{X_i}^l(v)$;

4. Generate the Green band by averaging the output of the previous step and scaling it by applying a gain and a bias: $\lceil \frac{Red+Blue}{2} \cdot 1.1 + 30 \rceil$. The gain and the bias inside the ceiling function aim at shifting the low to medium values of the Red and Blue bands that empirically appear to correspond in many cases to green areas, to higher levels of the green scale;
5. Crop the images in tiles of size $12,288 \times 12,288$ pixels: this operation is required in order to manipulate only the overlapped areas. The specific tile size was considered suitable in terms of I/O operations and file storage, while keeping in the same time enough information (samples) from the source image.

4.2.2. Tiles merging and rendering

This operation concerns the merging of the overlapped tiles. At tile level, an ordered list of images is being generated having as criterion the data domain size (in descending order). The first chosen tile (T_1 with data domain D_1) constitutes the canvas upon which the remaining tiles will be positioned. Next, the second tile (T_2 with data domain D_2) in the list is being read. Then, if $D_1 \cap D_2 = \emptyset \Rightarrow T_1 = T_1 \cup T_2$. Otherwise, the euclidean distance transform (DT) is computed over the binary array $D_c = D_1 \cap D_2$. The updated tile is composed as a weighted sum of the normalized DT , i.e. $T_1(D_c) = T_1(D_c) \odot (1 - DT(D_c) / \max(DT(D_c))) + T_2(D_c) \odot DT(D_c) / \max(DT(D_c))$, where the operator \odot signifies the pixel-wise multiplication. For the domain $D_n = \neg D_1 \cap D_2$, it holds $T_1(D_n) = T_2(D_n)$. Having only two tiles to process in the memory, this progressive operation turns to perform efficiently in terms of high-throughput computing, whilst allowing the incorporation of potentially new tiles without the need of re-processing big parts of the mosaic.

5. RESULTS AND PERFORMANCE METRICS

The algorithm fits the way a high-throughput computing cluster operates by allowing task parallelization, modularity and efficient I/O handling. Fig.1 demonstrates the scalability of JEODPP when handling the SITBX with both multithreading and single thread options; the capacity is measured on how much input data can be read, processed and stored back to EOS per second. Fig.2 shows the total elapsed time for both processes of false color composition (652 concurrent jobs) and tiles merging (800 concurrent jobs). It is worth

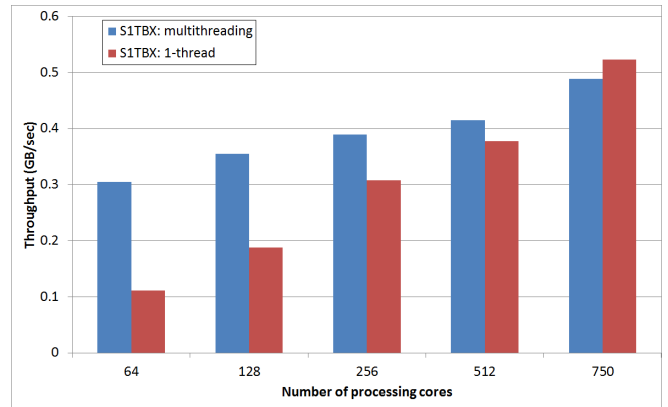


Fig. 1. The JEODPP scalability while running the SITBX with multithreading enabled (left bars) and by setting 1 thread per core with the option `-q` of the `gpt` command (right bars).

mentioning that both processes can execute simultaneously on the cluster by setting appropriate job priorities, shortening thereby significantly the total elapsed time.

The main challenges regarding the tiles mosaicking with the aim to produce an homogeneous result were the seasonality and the dissimilar orbit direction as steady effects, and occasionally, the distinct values among the different areas of the same scene due to suboptimal de-bursting and merging of the sub-swaths, the presence of artifacts, the signal saturation and the inadequate operation of thermal noise removal. Fig. 3 displays four such indicative cases (products prefix: S1A_IW_GRDH_1SDV). In the context of big data framework, exploiting all the available data covering a particular geographic area in a specific time span can lead to a better result due to the signal stability that inherently comes with the data redundancy (law of large numbers under the relaxed assumption of identically distributed measurements); this objective was out of the scope at the time of implementation.

An overview of the global S1-mosaic in the JEODPP interactive visualization is shown in Fig. 4. It is worth noting

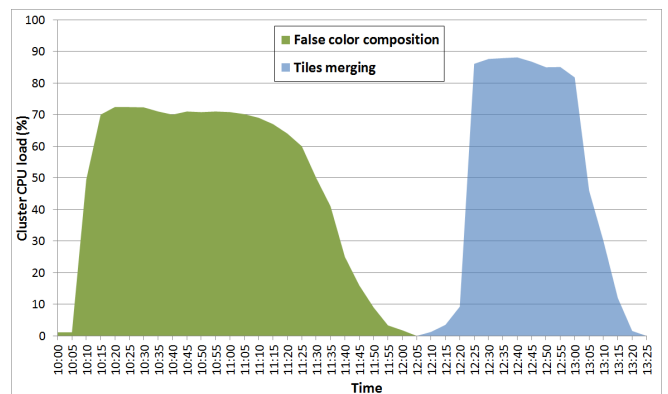


Fig. 2. The total execution time for the two stages of the main processing displayed against the cluster CPU load in a total of 912 available processing cores.

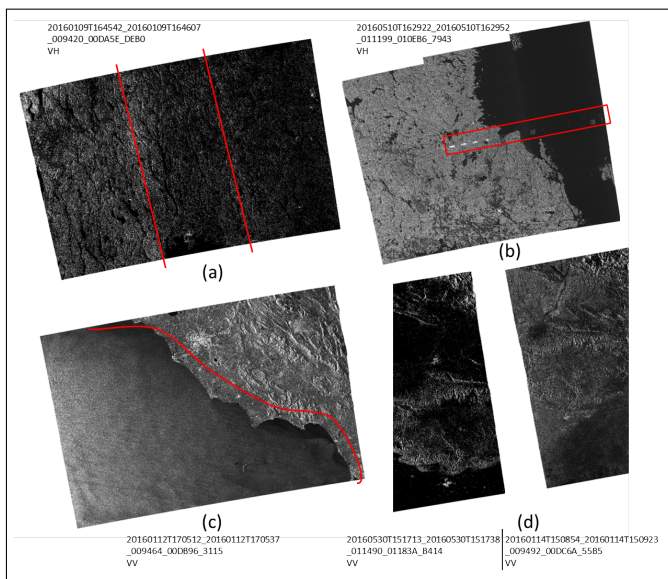


Fig. 3. Irregular cases on the available imagery: (a) intense discrepancies between adjacent sub-swaths; (b) artifacts; (c) signal saturation; (d) seasonality effect.

that no ancillary data like water mask or land cover layers have been used; the outcome has been produced based on the selected S1 data only.

Fig. 5 shows a snapshot of the S1-mosaic overlaid by the derived GHSL product (red pixels). Observing the left image, built-up in white-cyan contrasts vividly with the greenish and brownish lowlands, as well as with the dark color of the lakes.

6. CONCLUSION AND FUTURE DEVELOPMENTS

We presented an algorithmic workflow for mosaicking Sentinel-1 images using a false color composition such that a built-up layer derived from these images can be contrasted well, allowing the user to identify easily strange or unexpected effects. Statistical analysis on the available data provides

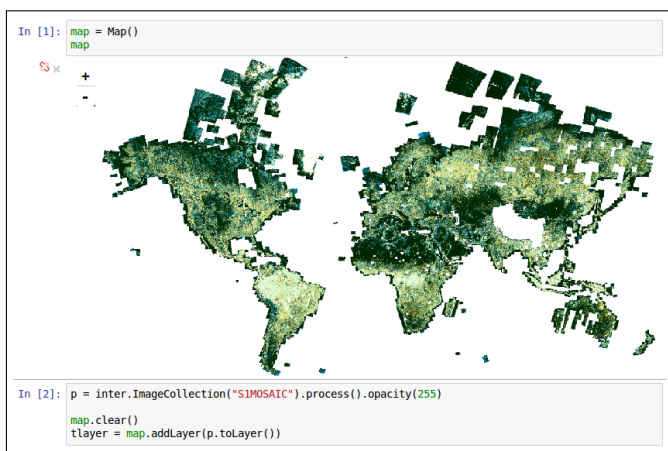


Fig. 4. The global mosaic from Copernicus Sentinel-1A data (EPSG:3857) with 19.11 spatial resolution.

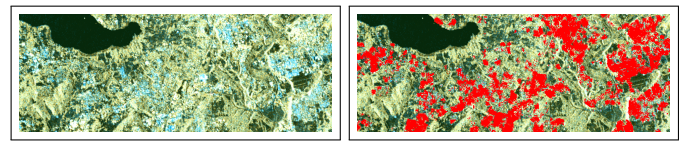


Fig. 5. Left: Example of S1-mosaic with built-up areas visible in bright blue; Right: Automatically extracted built-up areas from the GHSL displayed in red overlaying the S1-mosaic.

estimations that drive the discretization of the floating type images. Subsequently, the algorithm is executed in an automatic, nonparametric and progressive mode. As follow-up activity, we intend to focus on the optimization of the S1 product selection and on testing/adjusting on the fly false colour compositions through the interactive visualization.

7. REFERENCES

- [1] A first look at asia with sentinel-1a satellite imagery. <http://irri.org/s1a-mapping>. Accessed: 2017-06-30.
- [2] What is Copernicus? <http://www.copernicus.eu/main/overview>. Accessed: 2017-06-30.
- [3] Big data for the environment. http://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-11128/19488_read-47013/. Accessed: 2017-06-30.
- [4] Image: Dutch mosaic from copernicus sentinel data. <https://phys.org/news/2015-12-image-dutch-mosaic-copernicus-sentinel.html>. Accessed: 2017-06-30.
- [5] Sentinel-1A mosaic of Europe. www.esa.int/spaceinimages/Images/2015/10/Sentinel-1A_mosaic_of_Europe. Accessed: 2017-06-30.
- [6] Copernicus services data hub. <https://cophub.copernicus.eu/>. Accessed: 2017-06-30.
- [7] Global human settlement layer. <http://ghsl.jrc.ec.europa.eu/>. Accessed: 2017-06-30.
- [8] HTCondor. <https://research.cs.wisc.edu/htcondor/publications.html>. Accessed: 2017-06-30.
- [9] G. Adde, et al. Latest evolution of EOS filesystem. *Journal of Physics: Conference Series*, 608, 2015. doi: 10.1088/1742-6596/608/1/012009.
- [10] Corbane C., et al. Mass Processing of SENTINEL-1 and LANDSAT Data for Mapping Human Settlements at Global Level. In *Proc. of the 2017 conference on Big Data from Space (BiDS'17)*, 2017.
- [11] D. Merkel. Docker: Lightweight Linux containers for consistent development and deployment. *Linux J.*, 2014 (239), March 2014.
- [12] Soille P., et al. Towards a JRC Earth Observation Data and Processing Platform. In *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*, 2016. doi: 10.2788/854791.
- [13] Soille P., et al. The JRC Earth Observation Data and Processing Platform. In *Proc. of the 2017 conference on Big Data from Space (BiDS'17)*, 2017.

THE JRC EARTH OBSERVATION DATA AND PROCESSING PLATFORM

P. Soille, A. Burger, D. De Marchi, P. Hasenohr, P. Kempeneers, D. Rodriguez, V. Syrris, V. Vasilev

European Commission, Joint Research Centre (JRC)

Directorate I. Competences, Unit I.3 Text and Data Mining, via Fermi 2749, 21027 Ispra (VA), Italy

ABSTRACT

The JRC Earth Observation Data and Processing Platform (JEODPP) is a versatile petabyte-scale platform that serves the needs of a wide variety of projects. This is achieved by providing a cluster environment for batch processing, a web-based remote desktop access with a variety of software suites, and a web-based interactive visualisation and analysis ecosystem. These three layers are complementary and are all relying on a common hardware layer where the data is co-located with the processing services. The versatility of the platform is illustrated by a series of applications running on the JEO D P P.

Index Terms— EOS, Docker, Jupyter, HTCondor, batch processing, interactive visualisation, deferred processing

1. INTRODUCTION

Earth Observation is truly undergoing a big data shift following the free, full, and open availability of the data generated by the EU Copernicus programme and other initiatives. This shift motivated the development of the concept of the JRC Earth Observation Data and Processing Platform (JEODPP) to fulfill the needs of the JRC projects relying on geospatial data analysis in the context of their policy support activities [17]. The JEO D P P needs to address the needs of users with very different requirements originating from domain experts that developed methods and algorithms in a variety of programming languages over the years to occasional users with little or no knowledge of programming. The JEO D P P is a versatile platform that meets these requirements by following a multi-layer architecture where each layer serves the needs of specific user groups [16].

This paper summarises the main components of the JEO D P P. Section 2 presents an overview of the platform architecture and details its hardware layer as well as the chosen distributed file system. The different types of data stored on the JEO D P P are presented in Sec. 3. The main software layers are detailed in Sec. 4. Before concluding, an application gallery is presented in Sec. 5.

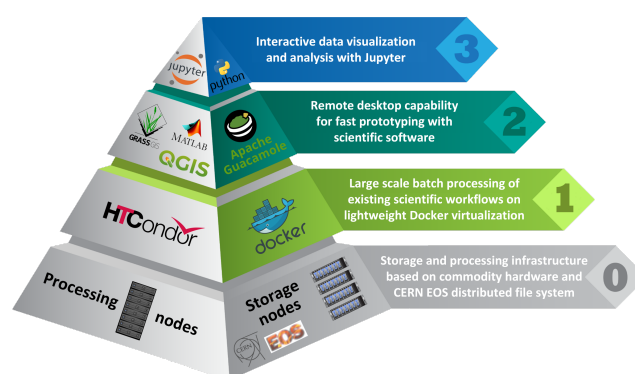


Fig. 1: The JEO D P P architecture: conceptual representation in the form of a 4-layer pyramid.

2. ARCHITECTURE

A conceptual representation of the JEO D P P architecture is sketched in Fig. 1 in the form of a four layer pyramid: the base layer (0) serves as a basis for the batch processing (1), the remote desktop (2), and interactive computational layers (3)

The base layer consists of scalable commodity hardware for processing and storage servers with directly attached storage (Just a Bunch of Disks or JBODs). Currently it is equipped with 16 storage servers for a gross capacity of 1.8 petabyte and 37 processing servers for a total of 992 core CPUs. The I/O bottleneck typically observed with network attached storage is avoided by considering appropriate high speed inter-communication topology. A key component behind the storage system is the underlying distributed file system that enables each processing server to have a unified view of all the files stored in the various disks attached to the storage servers. The JEO D P P relies on the EOS file system [1] developed by CERN to achieve this goal. EOS is mainly focused on low latency, high availability, ease of operation and low total cost of ownership. It is in production at CERN since 2011 and is managing more than 140 petabytes of raw disk space as of 2015 [12]. It allows for a flexible management of replica (redundancy) levels and works well in mixed hardware configurations. Files are accessed via the

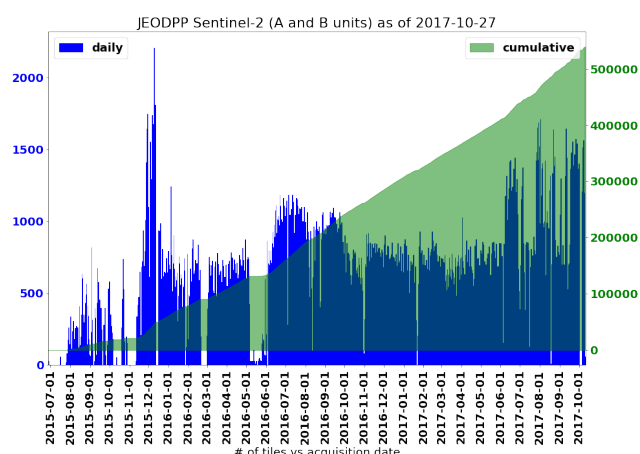


Fig. 2: Number of Sentinel-2 image tiles available on the JEODPP versus acquisition date.

Filesystem in Userspace (FUSE) client. This client acts as a translation layer between a POSIX compliant file system and the native XRootD protocol [4]. This means that all files on EOS storage can be accessed as if they were mounted on a single network file system volume.

3. DATA HOLDINGS

The JEODPP data holdings are driven by JRC user requests. The bulk of the storage space is currently used by Sentinel-1, Sentinel-2, and Landsat data. The data are downloaded from the various data hubs on a daily basis and the data catalogues are updated accordingly. For example, Figure 2 shows the number of Sentinel-2 tiles available on the JEODPP versus their acquisition date. The data are downloaded mainly for Europe and the tropical belt. They correspond to about a fourth of the total amount of available Sentinel-2 data (A and B units). In addition, all Sentinel-2 quicklooks are downloaded to enable application dependent optimal data selection, see example in [7]. Besides Sentinel data, other input data of interest to JRC projects such as Landsat Global Land Survey collections, Meteosat data, and VHR data over selected areas are also included. Further available raster datasets include products such as Global Human Settlement Layers [11], Global Surface Water layers [10], a series of Copernicus Land products as well as several base data such as Digital Elevation Models (EUDEM, 1 and 3 arsec SRTM, ASTER-DEM, etc.), the Copernicus CORE3 2.5m mosaic of Europe [15], a Sentinel-1 global mosaic [18], etc.

Contrary to some geospatial data cube representations [8] where all data sources are converted to a pre-defined coordinate reference system and grid followed by a fixed tiling and stacking along the time dimension, the raster data are stored on the JEODPP in the form of flat files as downloaded from the respective data sources. While pre-computed data cubes

provide a faster access to the values of the predefined grid cells along the time dimension, the flat file representation was preferred for its better suitability for general purpose analysis. Indeed, it avoids hard choices regarding the irreversible transformation of the input data to a fixed data cube representation and it is also suitable for heterogeneous data sources including vector data sets without the need to rasterize them.

Besides raster data, the JEODPP also holds a series of vector datasets like European and global administrative units, NATURA 2000 protected areas, EFFIS burnt-area time series, transport networks, etc.

4. JEODPP PROCESSING LAYERS

The three main processing layers are briefly described hereafter.

4.1. Batch processing and containerisation

Scientific workflows developed over the years can be applied to large image collections by distributing the workload to a series of processing nodes. Among the many available workload managers, HTCondor was selected because it is particularly suitable for applications where numerous independent jobs run in parallel without the need for inter-process communication, i.e., High-Throughput Computing (HTC). This is largely the case with satellite images that are often processed in parallel.

To eliminate the problem of installing and administrating application dependent software packages and libraries on the processing nodes and also to avoid the problem of applications having conflicting library requirements, the JEODPP makes heavy use of the light-weight virtualisation based on Docker containerisation [9]. Docker provides an operating-system-level virtualisation for flexible management of hardware resources and processing environments by allowing the existence of multiple isolated user-space instances called containers. In addition, the Docker container-based virtualisation technology integrates smoothly with the HTCondor task scheduler thanks to the so-called HTCondor Docker universe. For applications requiring inter-process communication using for example message passing interface (MPI), HTCondor can be used to submit jobs requiring multiple nodes while the job itself consists of a pool of Docker containers (one per node) managed by Docker SWARM.

4.2. Remote desktop

Some users have developed over the years applications based on dedicated software such as Matlab. Thanks to run-time execution environments, the processing can be launched through batch processing. However, this does not respond to the need for visualising the input or output data in the

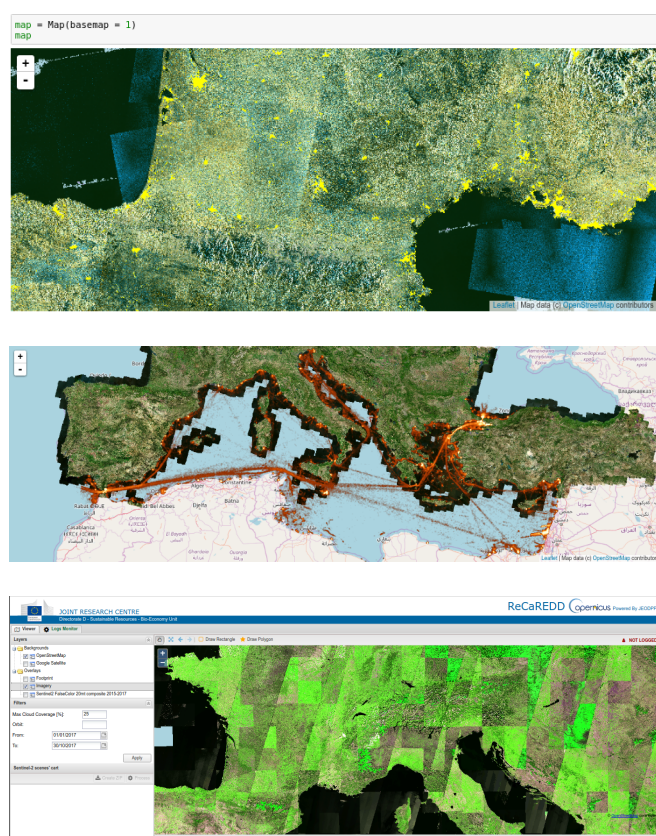


Fig. 3: JEODPP application gallery. Top: Global Human Settlement Layer [2] over Sentinel-1 mosaic. Middle: Ship detection from 2 years of Sentinel-1 imagery over the Mediterranean sea [13]. Bottom: Forest observatory [14].

development environment (besides the interactive visualisation capabilities detailed in Sec. 4.3). This is addressed by offering access to a web-based desktop environment based on Apache Guacamole, a web application that supports graphical access via remote desktop protocols directly in the browser based on HTML5, without the need for additional plugins. Various software libraries and tools such as GRASS, QGIS, and R used by the different JRC projects are provided.

4.3. Interactive visualisation and analysis

The JEODPP also offers the possibility to interact directly with the image data. This is achieved through a web interface that triggers the launch of a Jupyter Python notebook. A dedicated C++ library with Python bindings developed by the project offers the possibility to select and filter image collections and vector data sets, apply a series of processing steps, and render the resulting outputs in a map view area [3, 16]. The processing associated with the rendering is deferred in the sense that it only occurs when the tiles covering the map view area are requested, similarly to the approach followed by the Google Earth Engine [5]. Available transformations

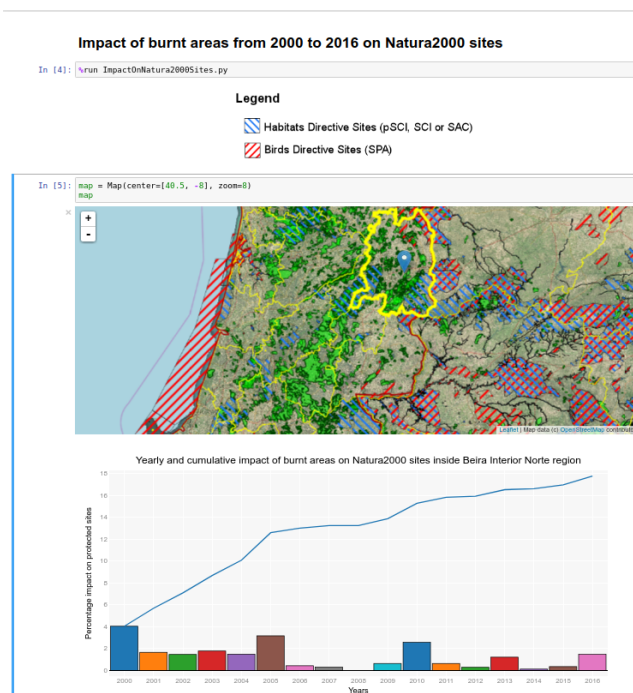


Fig. 4: Example of JEODPP Jupyter notebook showing the interactive computation of the NATURA2000 surface areas affected by forest fires over time within a user-selected territorial unit (a NUTS-3 region in Portugal in this example).

range from simple band combinations to complex connected component based segmentation. Besides users with Python programming skills, targeted information can be conveyed to non-technical stakeholders by simply providing an URL running a notebook with the Python code replaced by widgets, see example with the rendering of digital elevation data in [3].

5. APPLICATION GALLERY

Examples of applications running on the JEODPP are illustrated in Fig. 3: Global Human Settlement Layer [2], the Sentinel-2 web platform for browsing and processing Sentinel-2 imagery for forest cover monitoring over the tropics [14], and Sentinel-1 ship detection [13] using the JRC open-source SUMO software [6]. Besides the interactive visualisation, the Global Human Settlement Layer and ship detection applications are also relying on the batch processing layer for the massive computations involved. A Jupyter notebook with the interactive computation of the impact forest fires on NATURA 2000 sites is shown in Fig. 4.

6. CONCLUDING REMARKS AND OUTLOOK

Data-intensive computing for information retrieval from big geospatial data has recently emerged as a very active field given the availability of massive amounts of free and open

geospatial data. The proposed petabyte-scale platform enables the information extraction from large image data sets by users originating from different application domains with their specific data and software requirements. In addition, it contributes largely to knowledge sharing and collaborative working among users with very different levels of computer skills. The project is currently extending to other data sources such as social sensing in collaboration with the activities of the European Media Monitoring¹ by exploiting the geolocation data associated with the collected news and social media items.

7. REFERENCES

- [1] Adde, G. et al. “Latest evolution of EOS filesystem”. *Journal of Physics: Conference Series* 608 (2015). DOI: 10.1088/1742-6596/608/1/012009.
- [2] Corbane, C. et al. *Global Mapping of Human settlements with Sentinel-1 and Sentinel-2 data: Recent developments in the Global Human Settlement Layer*. Slides of presentation at WorldCover’2017, ESA, Frascati, Italy. Mar. 2017. URL: <http://worldcover2017.esa.int/files/2.2-p1.pdf>.
- [3] De Marchi, D., Burger, A., Kempeneers, P., and Soille, P. “Interactive visualisation and analysis of geospatial data with Jupyter”. In: *Proc. of the BiDS’17*. 2017, pp. 71–74.
- [4] Dorigo, A., Elmer, P., Furano, F., and Hanushevsky, A. “XRootD — A highly scalable architecture for data access”. *WSEAS Transactions on Computer Science* 4.4 (Apr. 2005), pp. 348–353. URL: http://www.researchgate.net/publication/234817900_XROOTDTXNetFile_a_highly_scalable_architecture_for_data_access_in_the_ROOT_environment.
- [5] Gorelick, N. et al. “Google Earth Engine: Planetary-scale geospatial analysis for everyone”. *Remote Sensing of Environment* (2017). DOI: 10.1016/j.rse.2017.06.031.
- [6] Greidanus, H., Thoorens, F.-X., Kourti, N., and Argentieri, P. “The SUMO Ship Detector Algorithm for Satellite Radar Images”. *Remote Sensing* 9.3 (2017), p. 246. DOI: 10.3390/rs9030246.
- [7] Kempeneers, P. and Soille, P. “Optimising Sentinel-2 image selection in a big data context”. In: *Proc. of the BiDS’17*. 2017, pp. 177–180.
- [8] Lewis, A. et al. “The Australian Geoscience Data Cube: Foundations and lessons learned”. *Remote Sensing of Environment* (2017). DOI: 10.1016/j.rse.2017.03.015.
- [9] Merkel, D. “Docker: Lightweight Linux Containers for Consistent Development and Deployment”. *Linux J.* 2014.239 (Mar. 2014). URL: <http://dl.acm.org/citation.cfm?id=2600239.2600241>.
- [10] Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. “High-resolution mapping of global surface water and its long-term changes”. *Nature* 540.7633 (2016), pp. 418–422. DOI: 10.1038/nature20584.
- [11] Pesaresi, M. et al. *Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014*. Tech. rep. EUR 27741 EN. Joint Research Centre of the European Commission, 2016. DOI: 10.2788/253582.
- [12] Peters, A., Sindrilaru, E., and Adde, G. “EOS as the present and future solution for data storage at CERN”. *Journal of Physics: Conference Series* 664 (2015). DOI: 10.1088/1742-6596/664/4/042042.
- [13] Santamaria, C. et al. “Mass processing of Sentinel-1 images for maritime surveillance”. *Remote Sensing* 9.7 (2017), pp. 678/1–678/20. DOI: 10.3390/rs9070678.
- [14] Simonetti, D. et al. *Sentinel-2 Web platform for REDD+ monitoring. Online web platform for browsing and processing Sentinel-2 data for forest cover monitoring over the Tropics*. JRC Technical Report. Joint Research Centre of the European Commission, 2017. DOI: 10.2760/790249.
- [15] Soille, P. “Seamless Mosaicing of Very High Resolution Satellite Data at Continental Scale”. In: *Proc. of the 2014 Conference on Big Data from Space (BiDS’14)*. Nov. 2014, pp. 222–223. DOI: 10.2788/1823. URL: <http://publications.jrc.ec.europa.eu/repository/bitstream/JRC92135/soille2014bids.pdf>.
- [16] Soille, P. et al. “A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data”. *Future Generation of Computer Systems* (2017). DOI: 10.1016/j.future.2017.11.007.
- [17] Soille, P. et al. “Towards a JRC Earth Observation Data and Processing Platform”. In: *Proc. of the BiDS’16*. 2016, pp. 65–68. URL: <http://publications.jrc.ec.europa.eu/repository/bitstream/JRC98089/soille-et-al2016bids.pdf>.
- [18] Syrris, V., Corbane, C., and Soille, P. “A global mosaic from Copernicus Sentinel-1 data”. In: *Proc. of the BiDS’17*. 2017, pp. 268–271.

¹EMM: <http://newsbrief.eu>.

A PLATFORM FOR MANAGEMENT AND EXPLOITATION OF BIG GEOSPATIAL DATA IN THE SPACE AND SECURITY DOMAIN

Sergio Albani, Michele Lazzarini, Paulo Nunes, Emanuele Angiuli

European Union Satellite Centre, Apdo de Correos 511, 28850 Torrejón de Ardoz, Spain

ABSTRACT

Building on the collection of user requirements from a number of key stakeholders in the Space and Security domain and on the experience gained in several H2020 projects, the SatCen RTDI Unit developed a platform providing the possibility to access, process, analyse and visualise satellite and collateral data. The platform is the first result of the efforts currently performed by SatCen in implementing Big Data and Cloud Computing paradigms for the processing of geospatial data.

The services currently available on the platform allow for the discovery and exploitation of Earth Observation data (in particular Sentinel-1 and Sentinel-2) as well as the visualization of social media data from open sources as Twitter.

Index Terms - Earth Observation, Big Data, Space and Security, Sentinel, Platform

1. INTRODUCTION

Within its Research, Technology Development and Innovation (RTDI) activities, the European Union Satellite Centre (SatCen) is implementing new solutions with regard to the whole data lifecycle to deal with the main challenges in the Space and Security domain.

A number of internal and external initiatives are taking place (e.g. the participation in H2020 projects such as BigDataEurope¹, EVER-EST², NextGEOSS³ and BETTER) to collect user requirements from key stakeholders in the Space and Security domain and evaluate new trending technologies. The key challenge is to improve the capability to handle a huge Volume of data produced with high Velocity by a Variety of sources ensuring Veracity and producing Value.

It has been assessed that technologies and techniques as Big Data, Cloud Computing and Machine Learning can increase the efficiency of accessing, processing, analysing

and visualising Earth Observation (EO) data as well as geotagged collateral information; moreover, automatic services (mainly based on open source tools) can facilitate the whole data management chain [1]. Therefore, the SatCen RTDI Unit developed a Service Oriented platform making use of Big Data and Cloud Computing technologies to manage and exploit Big Geospatial Data.

The interest of the Space and Security domain stakeholders in monitoring or assessing the situation on specific locations was the rationale for designing the platform considering the Area of Interest (AoI) as the core element of the system. This “AoI-centric” approach allows the user to define an AoI, to launch (accordingly to the available datasets) the service to run on it and to visualise relevant operational information.

2. DATA

Earth Observation data are currently showing an unprecedented scenario in terms of Volume (only Sentinel-1 and Sentinel-2 missions have produced more than 4.58 TB per day in 2016⁴), Variety (data are coming from different sensors in orbit on several governmental and commercial satellites), Velocity (data have to be received and processed in a short time frame to allow the provision of 24/7 information to users requiring fast responses), Veracity (decision making and operations require reliable sources) and Value (information to be provided has to be useful and clear).

Therefore it is crucial to improve the capacity to access and analyse such huge amount of complex data in order to timely provide decision-makers with clear and useful information.

In the development of the RTDI Platform, specific attention has been dedicated to guarantee access and usage of open data such as the ones from Sentinel missions. The Sentinel missions are operated by the European Space Agency (ESA) in the framework of the Copernicus programme funded and managed by the European Commission.

¹ <https://www.big-data-europe.eu/>

² <http://ever-est.eu/>

³ <http://nextgeoss.eu/>

⁴ <https://earth.esa.int/documents/247904/2955773/Sentinel-Data-Access-Annual-Report-2016>

The access and use of Copernicus Sentinel Data and Service Information is regulated under EU law [2, 3]. The free, full and open data policy adopted for the Copernicus programme foresees access available to all users for the Sentinel data products, via a simple pre-registration on the Copernicus Open Access Hub (COAHub, formerly known as SciHub) or other Sentinel collaborative ground segments. Currently other hubs, as the International Hub (IntHub) and the Copernicus Services Hub (ServHub), are open for specific entities.

Leveraging the value of open, large scale and continuous data provision, the RTDI Platform is currently able to work with datasets coming from Sentinel-1 and Sentinel-2 missions.

Sentinel-1A and -1B satellites were launched in 2014 and 2016 respectively, carrying a C-band Synthetic Aperture Radar [4] while Sentinel-2A and -2B satellites were launched in 2015 and 2017 respectively, carrying a Multi Spectral Instrument [5], an optical sensor able to acquire data in VIS, NIR and SWIR. Full-resolution (around 10 meters) Sentinel-1 Level-1 GRD images are at this time available in the platform for processing services (i.e. *Change Detection* and *Continuous Monitoring*), while GRD, SLC and OCN are available for the *Search & Download* service. Sentinel-2 Level-1C data (at 10, 20 and 60 m) are at this time available in the platform for the *Search & Download* and the *Indices* services.

The Medium-High Resolution data from Sentinel missions can provide an added value to the current image interpretation methodologies, based mainly on Very High Resolution data, used by the Space and Security community (e.g. Sentinel-1 and Sentinel-2 data can be used for monitoring of large areas, for supporting continuous monitoring tasks and for pro-actively monitoring potential Aols in view of possible tasks).

With regard to collateral data, the current focus is on social media data from Twitter. Tweets are available as plain text (Twitter messages) along with metadata in JSON format, searchable via a free Twitter Public Streams API. These data are accessible in the platform through the *Social Sensing* service.

3. PLATFORM ARCHITECTURE AND SERVICES

3.1. Architecture

The RTDI Platform high-level architecture is depicted in Figure 1.

The platform has been designed as a J2EE web application, i.e. a client-server software application where the client (or user interface) runs in a web browser and interacts with its server. The server has been implemented in Java and runs in a web server (or servlet container).

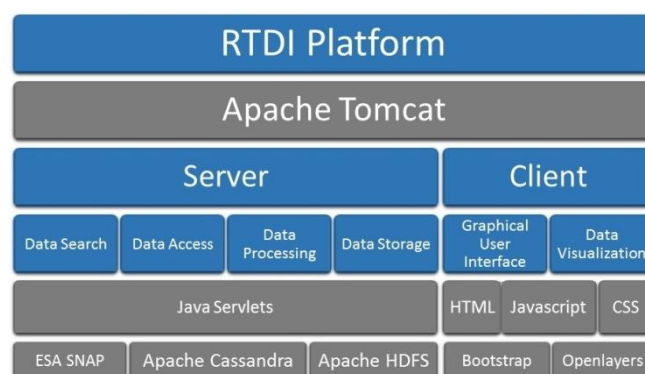


Figure 1. RTDI Platform high-level architecture

The platform design follows a Service Oriented Architecture⁵ design pattern, implemented by a HTTP RESTful⁶ API, exposing its services (*Search & Download*, *Change Detection*, *Continuous Monitoring*, *Indices* and *Social Sensing*) to be consumed by a WebGIS client. The implementation is based on open source technologies and languages like Java, Javascript, OpenLayers⁷, JQuery and AJAX⁸. The platform back-end runs on top of an Apache Tomcat server while its storage layer makes use of a NoSQL database (Cassandra⁹) and a distributed file system (Hadoop Distributed File System¹⁰). Processing services are mainly based on the ESA Sentinel Application Platform (SNAP¹¹) libraries, developed on a multi-threaded environment.

The RTDI Platform high-level architecture is composed by the following functional components.

Server

- Data Search implements the functionalities needed to search for satellite images metadata in the Sentinel ESA ServHub, according to user-defined criteria via the Graphical User Interface (e.g. AoI, timeframe, Sentinel mission);
- Data Access implements the functionalities needed to access (and download) user selected satellite images from the ServHub. The download is performed through a HTTPS connection and data are locally stored for further processing;
- Data Processing implements the Sentinel-1 satellite images' processing in which images are automatically pre-processed by chaining a set of SAR processing operators. The corresponding pre-processed products are then ready to be used in subsequent processing;

⁵ <https://www-01.ibm.com/software/solutions/soa/>

⁶ <https://codewords.recurse.com/issues/five/what-restful-actually-means>

⁷ <https://openlayers.org/>

⁸ https://www.w3schools.com/xml/ajax_intro.asp

⁹ <http://cassandra.apache.org/>

¹⁰ https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

¹¹ <http://step.esa.int/main/toolboxes/snap/>

- Data Storage implements and manages the local storage of all the data, in particular the downloaded satellite images and the corresponding pre-processed products to be (re)used in subsequent processing requests as well as all the information on the AoIs of each user.

Client

The client component has been designed in a decoupled way, providing the graphical means to access the server functional components previously described, and implementing the Data Visualization functional component. The client runs in a Web browser, while the server exposes a set of web services in the form of an API returning JSON to the client, which is then responsible to construct the web page(s), inserting the returned data into the page body (HTML) by means of DOM¹² manipulation. This task is performed via JavaScript, being the calls to the server made through AJAX. Every time the user interacts with the GUI (e.g. drawing an AoI, searching for data or requesting a processing), the browser performs one or more asynchronous HTTP requests to the server API fetching the data; the server retrieves the requested information and returns it to the browser; finally, the browser injects it into the web page body.

3.2. General Commands

The RTDI platform implements security access mechanisms (authentication/authorization) and the standard WebGIS functionalities (e.g. Basemap selection).

Being the platform “AoI-centric”, particular attention has been given to the study of the area identification: the search function is based on Geonames¹³ and when an AoI is created, tags can be associated in order to support a possible search for the *Social Sensing* service. All the AoIs created are stored in the user profile as well as the results of the services (e.g. maps of changes from the *Change Detection* service or specific maps from the *Indices* one), providing all the available information on a specific area from a single access point (Figure 2).

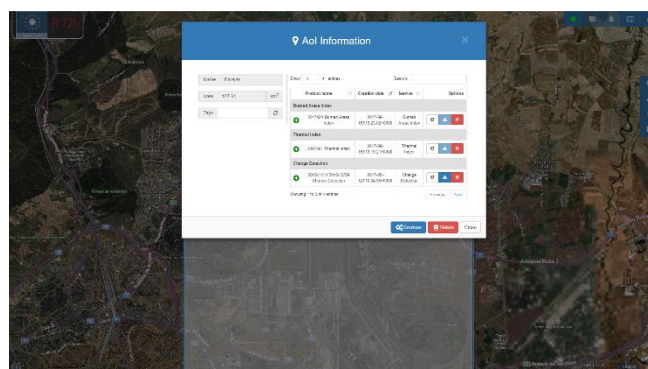


Figure 2. AoI panel with associated products

3.3. Services

Through the platform GUI (Figure 3), the user can currently execute on the selected AoI the following services:

- *Search & Download*;
- *Change Detection*;
- *Continuous Monitoring*;
- *Indices*;
- *Social Sensing*.

3.3.1. Search & Download

The *Search & Download* service allows selecting specific Sentinel-1 and Sentinel-2 images from the ServHub within a timeframe and downloading them.

The user can set several parameters for the data search (e.g. Product Type, Sensor Mode, Path Direction and Polarization for Sentinel-1, and Max Cloud Cover % for Sentinel-2) and download the image on its own machine.

3.3.2. Change Detection

The *Change Detection* service allows to select a pair of images from the Sentinel-1 archive, within a timeframe, and to identify clustered changes through suitable algorithms.

The service launches a set of chained processing modules based on SNAP, i.e. Subset, Orbit Correction, Thermal Noise Removal, Calibration and Terrain Correction. Successively, an in-house Change Detection algorithm identifies the areas with changes. The output of the Change Detection is a raster product containing the pixels where changes have been detected. In order to improve the visualization of the output, the changed pixels are aggregated in clusters by applying the DBScan [6] algorithm (creating polygons with minimum 2x2 pixels); results can be exported as KML¹⁴.

¹⁴ <https://developers.google.com/kml/>

¹² <https://www.w3.org/TR/DOM-Level-2-Core/introduction.html>

¹³ <http://www.geonames.org/>

3.3.3. Continuous Monitoring

The *Continuous Monitoring* service allows to select a starting image from the Sentinel-1 archive and to trigger an automatic Change Detection (with the processing chain used in the *Change Detection* service) every time a new image is available in the archive (given the same acquisition conditions, e.g. time and view angle).

With this service the user does not need to regularly check if a new Sentinel-1 image is available in the archive and activate the Change Detection manually as the system is able to perform the required operation automatically, storing the results in the user profile and sending a notification every time that a Change Detection is completed.

3.3.4. Indices

The *Indices* service allows to access images from the Sentinel-2 archive and visualise specific bands math operations in RGB or greyscale.

The service is based on the creation of a Web Map Service (WMS) from a specific set of predefined indices (e.g. True Colour, False Colour, Vegetation, Soil, Burned Areas): the user can either display the indices on the GUI or access the single bands and metadata of each Sentinel-2 scene located on the Amazon Web Service¹⁵.

3.3.5. Social Sensing

The Social Sensing service allows to search among Tweets on a set of per-user predefined Twitter accounts through specific keywords (a.k.a. Tags). The accounts have been limited to have a first screen of all web information; it is foreseen the possibility for the user to crawl only specific Twitter sources to personalize its search.

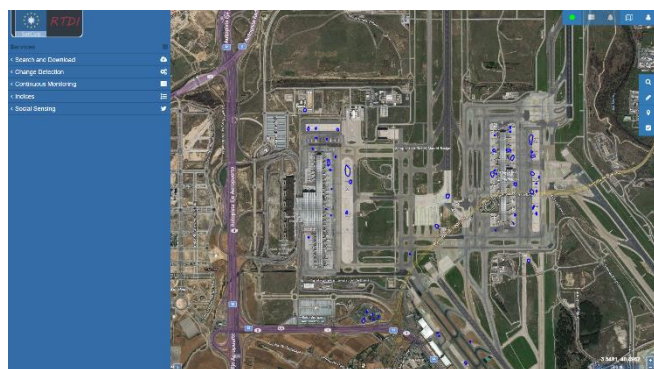


Figure 3. RTDI Platform GUI and services list

4. CONCLUSIONS

The RTDI Platform represents an innovative approach for an EO Service Oriented Platform, using the AoI as central point for the activation of the available services. Currently

¹⁵ <http://sentinel-pds.s3-website.eu-central-1.amazonaws.com/>

SatCen is actively experimenting the services provided by the platform, having already downloaded and processed large amounts of Sentinel data of interest.

The operational usage of the RTDI Platform has already demonstrated the possibility to save time with regard to the manual access, processing, analysis and visualisation of EO data and output products from a single access point.

In the future, the ingestion of additional data will be explored and further services (deployed internally or available via OGC standard interfaces) will be added to the platform, increasing the number of functionalities made available to the users.

5. REFERENCES

- [1] S. Albani, M. Lazzarini, M. Koubarakis, E.K. Taniskidou, G. Papadakis, V. Karkaletsis, and G. Giannakopoulos, "A pilot for Big Data exploitation in the Space and Security domain", Proceedings of the 2016 conference on Big Data from Space (BiDS'16), JRC Publications Office, pp. 196-200, 2016.
- [2] Commission Regulation (EU) No 4311 final of 12 July 2013 of the Council of the European Union supplementing Regulation (EU) No 911/2010 of the European Parliament and of the Council on the European Earth monitoring programme (GMES) by establishing registration and licensing conditions for GMES users and defining criteria for restricting access to GMES dedicated data and GMES service information.
- [3] Regulation (EU) No 377/2014 of the European Parliament and of the Council of 3 April 2014 establishing the Copernicus Programme and repealing Regulation (EU) No 911/2010.
- [4] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. Navas Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, and F. Rostan, "GMES Sentinel 1 Mission", *Remote Sensing of Environment*, 120, pp. 9-24, 2012.
- [5] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, P. Bargellini, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services", *Remote Sensing of Environment*, 120, pp. 25-36, 2012.
- [6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), pp. 226-231, 1996.

MUSCATE A VERSATILE DATA AND SERVICES INFRASTRUCTURE COMPATIBLE WITH PUBLIC CLOUD COMPUTING

Joëlle DONADIEU¹, Simon BAILLARIN¹, Marc LEROY¹, Robert NGO¹, Julien NOSAVAN¹, Arnaud SELLE¹, Celine L'HELGUEN¹
Roger RUTAKAZA MANENO², Thierry SEGUR², Bastien JULIE², Laurent FAVOT²

¹ CNES, 18, avenue Edouard Belin 31401 TOULOUSE CEDEX 4, France – joelle.donadieu@cnes.fr

² CAPGEMINI, 109 avenue du Général Eisenhower 31000 TOULOUSE, France - roger.rutakaza@capgemini.com

ABSTRACT

MUSCATE is the part of the THEIA data land centre dedicated to process optical images from SPOT, LANDSAT and SENTINEL-2 satellites. These images are intended for both scientific community and institutional actors.

MUSCATE (Multi-satellites, multi-sensors and multi-temporal THEIA data centre) is now operationally used on CNES Computing Centre to process and distribute up to 1600 products a day (<https://theia.cnes.fr>).

Beyond this objective, it has been envisioned to use and provide MUSCATE in the frame of wider collaborations and therefore to deploy the framework on other computing means and in particular Public Clouds.

After an overall presentation of the initial MUSCATE architecture designed to be used on CNES Computing Centre, the paper focuses on the improvements made in order to deploy the framework on Public Cloud together with the associated sizing and trade-offs (design/costs).

Finally, the paper presents the very promising results in term of performance and scalability of the solution once used on a cloud computing architecture.

Index Terms— THEIA, MUSCATE, Cloud Computing, SENTINEL-2, data valorization

1. INTRODUCTION

THEIA is a French national multi-agency organization which promotes the use of satellite data by scientific community and public policy actors. This consortium aims at helping monitoring human and climate impacts on ecosystems and territories by delivering a large panel of products and mutualised services allowing the user community to get the largest benefit of data and products from space missions.

As Data and Services Infrastructures of THEIA, MUSCATE [1] is designed to acquire, process and distribute automatically high resolution satellite images from SPOT 1 to 5, LANDSAT 5-7 and 8, and SENTINEL-2 in order to elaborate and distribute value added products (L1C, L2A, L3...) covering France territories and worldwide areas of interest.

CNES has developed with CAP GEMINI the MUSCATE framework and has deployed it on CNES High Performance Computing (HPC) centre based in Toulouse (France).

MUSCATE is directly linked to the French Sentinel products exploitation platform (called PEPS [6]) to download Sentinel-2 data and generate value-added products. PEPS relies on CNES (French National Space Agency) data storage facility HPSS (High Performance storage System) based in Toulouse (France).

In order to reach the objective of providing access to these value added products to a maximum number of users and partners (institutional, scientific...) covering new areas of interest from any point of the world, it has been envisioned to deploy MUSCATE on others computing means and in particular cloud infrastructure.

This document presents firstly MUSCATE architecture on CNES HPC centre, then improvements made in order to deploy the framework on Public Cloud together with the associated sizing and trade-offs (design/costs) and finally the first results in terms of performance and scalability of the solution once used on a cloud computing architecture.

2. MUSCATE OVERVIEW

MUSCATE Data and Services Infrastructure is based on acquisition module which collects products (SPOT 1 to 5, LANDSAT 5-7 and 8, and SENTINEL-2...) and catalogue component for data and product management. In addition, MUSCATE integrates existing CNES components whose

quality and efficiency have been proven within other satellite projects, in order to minimize development cost as well as to optimise computation time: PHOEBUS, SIGMA and MAJA. PHOEBUS is a workflow orchestrator: it runs processes over distributed resources, manages their priority and, via its interface, allows monitoring and control of their progress. It enables to define specific workflows for each type of acquired product. SIGMA integrates complex algorithms to correct the acquisition sensor model and to orthorectify satellite products. This orthorectification step ensures a correct multi-temporal co-registration of images [2]. Finally, MAJA implements a recursive and multi-temporal algorithm, designed from a combination of CESBIO and DLR algorithms [3] [4], which converts a level 1C product expressed in TOA reflectance in a level 2A product expressed in surface reflectance. It is then capable to process temporal series of images at high resolution, high revisit and under constant viewing angles like LANDSAT and SENTINEL-2 data.

New processing workflows are being currently integrated in the framework, such as snow cover products, multi-temporal synthesis products and other are already planned in the near future (for 2018 : bio-physical products, land cover maps, ...).

3. MUSCATE ARCHITECTURE

The architecture of MUSCATE is composed of two main software components, the Acquisition-Production Module and the Distribution Module.

The Acquisition-Production Module is in charge of data acquisition and processing and generates added value products (L1C, L2A, L3...) which are distributed by the “Distribution” Module to the scientist and public community.

The Acquisition-Production Module relies on CNES orchestrator PHOEBUS and on a “central registry” component called “Catalogue” which carries MUSCATE system intelligence.

PHOEBUS which is in charge of the processing orchestration sends jobs on the CNES High Processing Cluster and manages their priority. PHOEBUS enables to define specific workflows for each type of acquired products. Each workflow is broken down in steps which can be sent to the distributed resource manager (DRM) of the CNES HPC centre. PHOEBUS facilitates the integration of new algorithms in the overall architecture to process massive volume of data. Moreover it offers an integrated MMI which allows operators to follow production progress and to act if required. PHOEBUS also gives the possibility to regulate the number of jobs sent in parallel to the cluster.

The “Catalogue” component is consisting of a catalogue of image products together with a dedicated set of processing rules to automatically trigger the processing workflows once necessary input products are available (e.g. multi-temporal MAJA processing), on given areas of interests, for given temporal periods and with specific parameters (processing parameters, external data such as DEM, etc...).

Catalogue and PHOEBUS components are deployed on an independent server. They are based on Java technologies using Service Oriented Architecture on server side. The client side is built using Rich Client interfaces technologies (JavaFx and Swing).

On the other hand, users have a direct access to products in the Distribution Module in CNES system through MUSCATE research HMI. The “Distribution” Module uses web technologies (PHP, HTML, Javascript...). It is composed of two web sites using a VM (Virtual Machine) based on a shared infrastructure. These two web-sites are packed in CNES Apache&Tomcat software components CNES package called WEB-NG.

MUSCATE architecture has been designed to facilitate its integration on CNES HPC centre (cf. Fig. 1). This computing cluster has a large capacity which enables MUSCATE to process up to 200 Landsat, 600 Spot and 800 Sentinel-2 products a day.

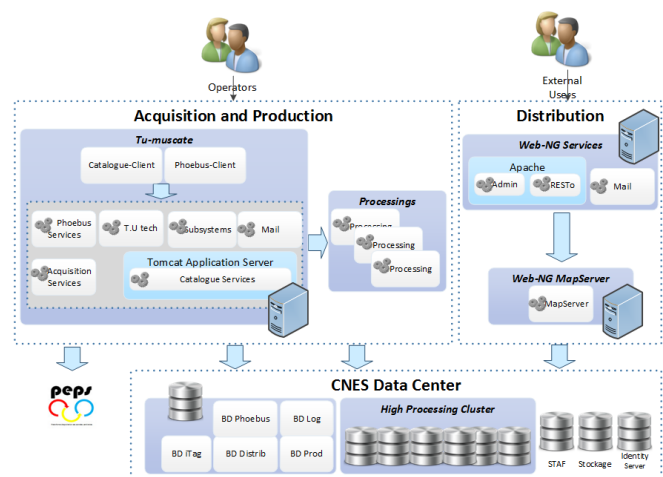


Fig 1 : Muscate architecture on CNES HPC centre

4. TOWARDS A CLOUD COMPATIBLE SOLUTION

A cloud infrastructure using public cloud has been envisioned. This choice takes its motivation in providing services to worldwide users without needing to provision or

integrate specific infrastructure and simplifying operation or maintenance tasks to be done on the MUSCATE system.

The Cloud solution also provides flexibility to deploy new algorithms and infrastructure scalability to deal with the evolution of the production needs. Thus, rules can be specified to automatically adjust the processing capabilities of the infrastructure according to the user needs.

To bring MUSCATE system Cloud compliant, several activities were defined and put in place.

Two main streams were defined related to the:

- Acquisition-Production module with:
 - Client side in order to allow access from any site,
 - Server side components in order to fit cloud constraints without any changes on the production version.
- Distribution module

4.1. Acquisition-Production

4.1.1. Client side

During design phase, the need to separate client side and server side was also identified in particular to allow remote access for operating the system. Regarding MMI, trade-off has been done between “ssh MMI” export from the cloud and deployment of client MMI on the remote PC. The second option has been finally choose to allow a performant access.

4.1.2. Server side

First, architecture deployment design has been revisited. The different components of the framework (catalogue and PHOEBUS) were deployed inside Virtual Machines (VM). This step allows being independent from the infrastructure and any cloud provider. In parallel the algorithms make use of computing VM being deployed inside DOCKER containers.

A dedicated acquisition plugin has been developed to download Sentinel-2 data from Amazon infrastructure.

Improvements have been made to optimize archive means.

4.2. Distribution

The “Distribution” component based on web technologies was the most simple to make cloud compliant. Nevertheless, for more flexibility and keeping cloud provider independence, the different software components are

deployed in Docker containers before their deployment on virtual machines hosted in Cloud infrastructure.

4.3. Target Architecture

In order to minimise the data transfer and benefiting from a previous project [5], a first deployment has been done on Amazon infrastructure to test Sentinel-2 L-2A processing.

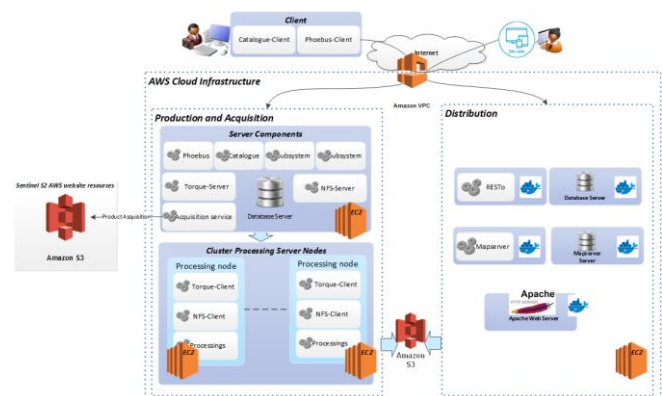


Fig 2: Target Muscate architecture

The infrastructure that supports the target architecture (described in Fig. 2) is composed of AWS EC2 instances in order to host the “Acquisition-Production” module and Processing nodes. One of the main advantages of using the cloud infrastructure is the scalability. Thus, depending on the processing performance requirements, new EC2 processing nodes could be added to the infrastructure transparently. The “Distribution” services are hosted on AWS EC2 instances too. The S3 instance is used, first of all, to share generated products between the “Acquisition-Production” and the “Distribution” modules and secondly to store products to be distributed. To keep control to the MUSCATE infrastructure, AWS VPC (Virtual Private Cloud) is used to isolate MUSCATE resources.

This deployment is going to be tested soon on other cloud providers (Orange, T-System...).

4.4. Performance and scalability

The performances obtained in this context were very promising. Comparing to the CNES HPC, an overhead of 15% was seen on AWS infrastructure with the similar configuration of MUSCATE on CNES HPC centre. This should be solved by using more powerful VM as processing nodes.

Beside the technical aspects, the cloud infrastructure allows optimising resources usage and cost. Thus, it deals with green IT because its business model is based on “Pay-per-use” principle, only consumed services are billed.

4.5. Business model

Associated business case model and cost studies per year were also evaluated in order to deal with the generation of 260 Sentinel2 products in a day equivalent to 5M.km².

To optimize the business model, it was decided to keep only three months of data.

For information, a summary of the business model is reported in Table 1. This includes:

- 1 VM for the Acquisition-Production,
- VM for the Processing Nodes,
- S3 instance for the storage,
- 1 VM for the Distribution,
- Product download.

Table 1. Summary of business model.

Services	Pricing \$	Resources
Production Server (EC2 instance)	2763	m1.xlarge(IR)
Processing Node (EC2 instance)	37563	m2.xlarge(IR)
Storage (S3)	9482	S3
Distribution Server (EC2 instance)	485	m3.medium
Product download	10535	
TOTAL	60828 \$	per year

5. CONCLUSION AND OUTLOOKS

As described, the scope of image processing embedded in MUSCATE is very wide and complex and will continue to evolve in the future to integrate new added value products.

Beyond its operational use on CNES HPC for THEIA, MUSCATE first results obtained on Cloud Computing infrastructure are very encouraging in term of computing performance as well as in term of scalability.

MUSCATE evolution towards Public Cloud computing opens to new opportunities in particular in the framework of collaborations (scientific, institutional, commercial...) and can therefore participate to the promotion and development of new Earth Observation products and services for a large number of users.

6. REFERENCES

- [1] J. Donadieu et al. “MUSCATE: Multi-satellites, multi-sensors and multi-temporal THEIA data centre”, Big Data From Space 2016, Tenerife, Spain.
- [2] S. Baillarin, P. Gigord and O. Hagolle, « Automatic Registration of Optical Images, a Stake for Future Missions: Application to Ortho-Rectification, Time Series and Mosaic Products », Geoscience and Remote Sensing Symposium, 2008, 2:II-1112-II-1115. doi:10.1109/IGARSS.2008.4779194, 2008
- [3] V. Lonjou, C. Desjardins, O. Hagolle, B. Petrucci, T. Tremas, M. Dejus, A. Makarau, S. Auer « MACCS-ATCOR joint algorithm (MAJA) », Proc. SPIE 10001, Remote Sensing of Clouds and the Atmosphere XXI, 1000107 (October 19, 2016)
- [4] O. Hagolle, M. Huc, D. Villa Pascual and G. Dedieu, « A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images », Remote Sensing of Environment 114 (8) (août 16): 1747-1755. doi:10.1016/j.rse.2010.03.002, 2010
- [5] S. Daniel, B. Koetz, T LeToan “GeoRice, Tech4Earth applications deployed on AWS”, Living Planet Symposium, ESA, May2016 (http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-1/Sentinel-1_sees_rice_paddy_drop_in_the_Mekong_Delta).
- [6] Plateforme d’Exploitation des Produits Sentinel <https://peps.cnes.fr>

COMBINING SMALL HOUSEKEEPING DATA LAKES INTO A SHARED BIG DATA INFRASTRUCTURE AT ESOC - ACHIEVEMENTS AND FUTURE EVOLUTION

Rui Santos^[1], Gustavo Marques^[2], James Eggleston^[1]

^[1] ESA/ESOC, Robert-Bosch Strasse 5, 64293 Darmstadt, Germany.

^[2] CGI, Rheinstrasse 95, 64295 Darmstadt, Germany

ABSTRACT

The ESA/ESOC Analysis and Reporting System (ARES) provides support for off-line storage, analysis and display of many types of operational data, such as TM packets, TM parameters and commanding. Combined with the fact that modern missions produce ever increasing volumes of data, has led to a significant expansion of requirements for the long term mission off-line data storage. This paper will demonstrate how a “Big Data” distributed cluster approach, based on a Hadoop ecosystem, was adopted at ESOC to overcome operational restrictions and technical limitations of previous solutions, as well as looking forward to future opportunities of new data analysis techniques and algorithms.

Index Terms— Big Data, Hadoop, ARES

1. INTRODUCTION

The ESA/ESOC Analysis and Reporting System (ARES) provides support for off-line storage, analysis and display of several types of operational house-keeping data, including: telemetry packet and parameter information; telecommand history and failure events; Spacecraft and Mission Control System events. ARES uses other ESA generic systems (such as the ESA Ground Operations Systems (EGOS) Data Dissemination System (EDDS[1]) and the EGOS User Desktop (EUD[2])).

However, off-line data warehouses usually manage data in engineering form to minimise dependencies from the processing systems, as well as, to maximise the data access performance. This design feature, combined with the fact that most modern missions produce ever increasing volumes of housekeeping data (e.g. improvement of on-board technology, the increase of space link rates, and mission complexity) has led to a significant growth of the level of requirements for the long term mission off-line data storage and performance, which have to be counterbalanced with the need for cost effective solutions.

Traditional implementation technologies, such as Relational Databases, are brought to the limit of their capabilities and missions are obliged to adopt strategies to alleviate pressure on the underlying technology (such as limiting the content and scope of the stored data). Nevertheless, even using these strategies to store off-line

operational house-keeping data, achieving linear performance scalability, considering the data volume, is difficult or impossible to achieve.

This can be contrasted with the “Big Data” approach, where such operational restrictions due to technological limitations are avoided, and many spontaneous user driven activities are simultaneously run and permanently available.

In the next sections we will report the operational usage and planned evolution of the ARES system regarding data storage, processing and analysis in a “Big Data” distributed Hadoop cluster approach based on HDFS, HBase, Yarn/MapReduce and Spark.

2. ARES DESIGN AND DATA FLOW

ARES is the most recent evolution of off-line data analysis systems at ESOC. It fulfils the same role and principles as the Mission Utility and Support Tools (MUST[3][4]) did in the last decade. It has been deployed operationally for mission such as Gaia, Exomars and Cluster.

Currently all data provided to ARES is still file based (binary, CSV, XML). In order to isolate it from mission specific changes and simplify the data import process, ARES is an EDDS client (see Fig. 1), relying on it to obtain the necessary data. This means ARES expects to receive and parse data already in an engineering format. Imposing this design restriction allows ARES to become a generic system that is capable of storing any type of parameter like data.

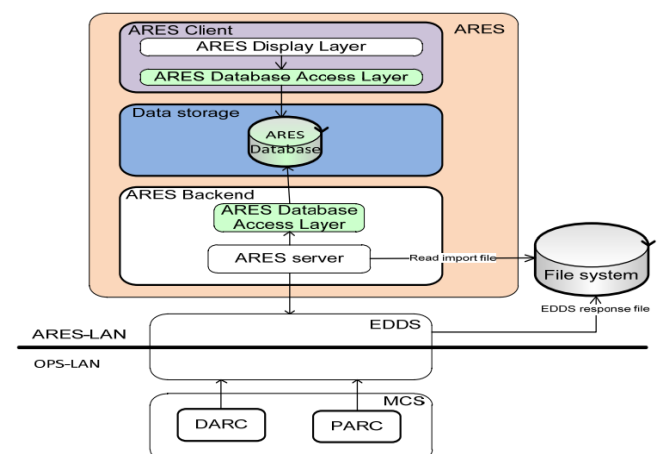


Fig. 1 – ARES simplified components overview

The traditional sources of data for ESOC missions are the Mission Control System (MCS) data archives (Packet Archive (PARC) and Data Archive (DARC) for TM parameter data).

Another key dependency of ARES is the ESA generic user interface framework, EGOS User Desktop (EUD), which provides several generic common displays, including TM parameter plotting capabilities (example provided in Fig. 2), which introduces the requirement of time range based retrieval of potentially long intervals.

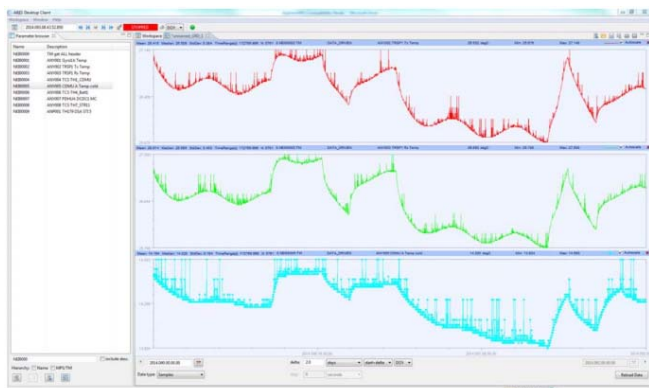


Fig. 2 – ARES graphical display

3. HADOOP ECOSYSTEM USED BY ARES

3.1. Introduction to the Hadoop ecosystem

Hadoop is an open source Apache licensed Java-based programming framework that supports the storage, mining and fast processing of large data sets in a distributed computing environment, while providing the ability to extend them with new derived parameters or generate entirely new data sets from existing data. As a highly distributed technology, Hadoop relies on a cluster of nodes which can be easily (horizontally) scalable.

ARES makes use of proven Hadoop functionality of services such as HDFS, Yarn/MapReduce, HBase, and Spark. The next sub-sections will provide a very brief introduction to each of these services.

3.1.1. HDFS

HDFS (which stands for Hadoop Distributed File System) is the backbone of the Hadoop framework. It has been designed to store very large data sets distributed across a cluster of nodes running in commodity hardware. HDFS support features like fault tolerance, scalability, data replication, and High Availability out of the box.

3.1.2. Yarn/MapReduce

Yarn/MapReduce is a software framework for resource management and writing applications which process potentially vast amounts of data in parallel on top of HDFS

and distributed across the cluster bringing high performance results. The framework facilitates the work distribution across the cluster by scheduling tasks, monitoring them, and re-executing any that have potentially failed.

3.1.3. HBase

Apache HBase is a column-oriented database that is usually integrated on top of HDFS, specifically designed and optimized for read performance while providing high level of scalability, reliability, and schema flexibility. HBase is also commonly integrated with Yarn/MapReduce.

3.1.4. Spark

Apache Spark is a fast execution engine supporting functional programming, acyclic data flow, in-memory calculations, streaming and complex analytics, embedded with fault tolerance.

3.2. Hadoop Use Case for ARES

Hadoop, with its inherent extensibility as a primary design goal, due to the market needs for cost effective solutions is able to deal with the large Terabyte and Petabyte level data sets, allow missions to extend the types of data stored in ARES and their correlations (such as e.g. storing all TM packet information and parameter values; something missions have repeatedly requested) without being concerned with performance degradation or suitability of the data storage system and backup over the mission lifetime, or in later years to provide cross mission analysis. As an example, given our estimation, a mission such as Gaia would require a volume of approximately 60TB for house-keeping TM data (packet and parameter) alone over 10 years. We anticipate future missions will have at least similar or more demanding requirements.

Currently, ARES receives and archives all operational house-keeping data in HDFS ensuring its veracity, integrity and isolation per mission. It is then processed by Yarn/MapReduce and hosted in HBase, not only for easy access, but also keeping it available for later analysis, through specific graphical interfaces (RCP, WEB) with multiple plotting and exporting capabilities. Also, at regular time intervals, statistical calculations are performed on existent data sets.

Using Spark, ARES also provides a generic extensible framework (currently based on Java and Scala) for deploying other algorithms, specifically for evaluation of any spacecraft telemetry data, enabling users to design and perform any data computing on the entire stored data set. With this, each mission now possesses the capability to define much more complex and imaginative data analysis operations, with minimal pre-required knowledge of the Hadoop technologies used under the hood, which can be easily plugged into the ARES infrastructure system. The framework requires several key inputs, such as parameters,

time periods and parameter aggregation calculations. Fig. 3 shows an example of the component diagram of the framework when used for a client such as DrMUST[5] (DrMUST is a data mining client that can support flight control engineers in their anomaly investigation tasks. It performs pattern matching and correlation analysis).

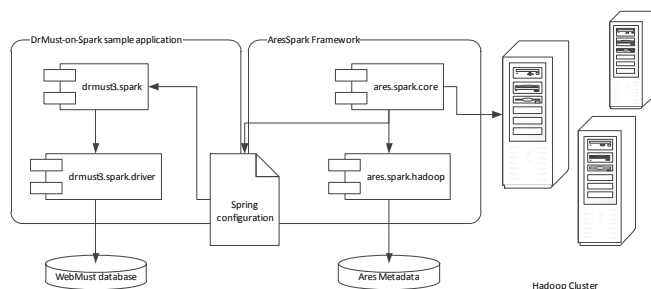


Fig. 3 – AresSpark Framework component diagram

This allows future data analysts, for instance principal investigators, flight control teams or flight dynamics systems to utilize the data through additional ad-hoc queries maximizing the value of the stored data, making use of modern machine learning algorithms and methods to tackle complex problems, such as space situational awareness.

3.3. ARES operational deployment at ESOC

Traditionally each ESOC mission would have their own self-controlled and managed ARES deployment. This would create a proliferation of relatively large data lakes, which in turn would have to be independently monitored and maintained during the lifetime of each mission. With the introduction of Hadoop, it was possible to centralize this support as part of the service catalogue provided by the ESOC IT department (storage and processing as a service type of approach).

Currently, the Hadoop production cluster hosts 3 concurrent missions and is composed of 8 machines (around 100TB capacity and 96 computing cores available to Yarn). The service capacity and usage is regularly evaluated so it can be scaled to meet the demands (based on current usage, the medium term plan is to add 2 nodes/15TB capacity each year). Due to this effortlessly horizontal scalability feature, Hadoop also allows ESOC missions to make use of it for the purpose of long term data preservation.

4. BENEFITS AND FUTURE PLANS

With ARES and Hadoop we have opened the door for new operational opportunities and benefits provided by supporting the “Big Data” paradigm. Not restricted to, but mainly including numerous possibilities as simplification of mission preparation (e.g. Bepicolombo, Aeolus) by making use of already existing storage and processing facility, scalable long-term archiving, data backup and replication, mission specific data mining and correlation analysis

algorithms, stream processing, machine learning, multi-mission application architecture. Fig. 4 shows a glimpse of the different evolution possibilities, that can include a variety of different user applications, types of data (including data currently not managed by ARES) and different ways to transverse and aggregate data at different stages of the computation.

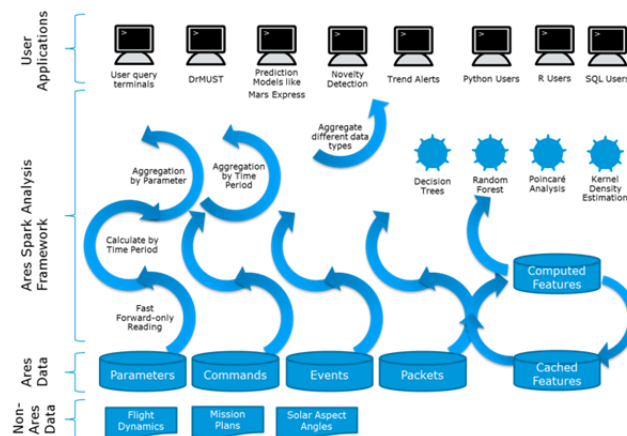


Fig. 4 – Evolution avenues

The side effect of such approach, besides the added complexity of managing all the Hadoop components and their moving parts, is the fact that Hadoop typically processes data in batch mode and expects the totality of the data set to be available when all the computation is executed, as well as outputting the entirety of the expected outcome upon completion. This lead us to the point where the main goal becomes to obtain results close to real-time. An alternative approach is the usage of low latency stream processing, whilst allowing the utilisation of machine learning algorithms and techniques, further improving the users experience, by providing them the means to monitor, not only currently available data, but also advanced correlations of near to real-time data. Spark already provides all the means to integrate and combine all these options in the same application, in very similar way presently done in our current batch processing approach, due its capability to read data from e.g. HDFS. Work is ongoing to prototype streaming solutions from different data sources (such as EDDS or EUD framework – through a generic interface).

ESA is modernising the mission control system infrastructure (as part of the European Ground System – Common Core EGS-CC/EGOS-CC[6] project). ARES is part of this evolution but through use of EDDS as the main source of data, ARES will remain mostly unaffected. EGS-CC is also designed to make use of Hadoop as foundation for his data storage needs and therefore can greatly benefit from the synergy and experience already provided by the work done for ARES. Part of the future work at ESOC is to extend the Hadoop support to reach other areas as required by EGOS-CC.

5. CONCLUSION

ARES is an ESA system to provide generic, yet powerful mechanisms to store and analyse house-keeping data. With the introduction of Hadoop as part of the ARES data storage component, we have evolved ARES into a more flexible system therefore delivering additional value to ESA missions operations. By removing the restriction on the amount and scope of data being stored, as well as, making a better use of computing resources, we are creating opportunities for future ESA missions to better understand problems and explore opportunities. The change from isolated mission data silos into a shared, distributed and common solution we simplify part of the deployment process for each mission and can better and more efficiently harmonise solutions.

6. REFERENCES

- [1] F. Flentge, R.Santos, "Ground Segment File Handling for Ground Station and Spacecraft Operations", *Proceedings of the 10th International Conference on Space Operations (SpaceOps)*, April 2010.
- [2] J. Schuetz, "EGOS User Desktop A Generic User Interface Framework for Ground Segment Software", *Proceedings of the Ground System Architectures Workshop (GSAW)*, February 2014.
- [3] J. Martinez-Heras, A. Baumgartner, A. Donati, "MUST: Mission Utility and Support Tools", *Proceedings DASIA 2005 conference*, 2005.
- [4] A. Baumgartner, J. Martinez-Heras, A. Donati, M. Quintana, "MUST – A Platform for Introducing Innovative Technologies in Operations", *Proceedings of the ISAIRAS 2005 conference*, 2005.
- [5] J. Martinez-Heras, A. Donati, B. Sousa, J. Fischer, "DrMUST – a Data Mining Approach for Anomaly Investigation ", *Proceedings of SpaceOps 2012 Conference*, June 11 – 15, 2012.
- [6] M. Pecchioli, A. Walsh, "The EGS-CC based Mission Control Infrastructure at ESOC", *Proceedings of SESP 2017 conference*, 2017.

EVOLVING JASMIN: HIGH PERFORMANCE ANALYSIS AND THE DATA DELUGE

Neil Massey¹, Philip Kershaw¹, Matt Pritchard¹, Matt Pryor¹, Sam Pepler¹, Jonathan Churchill² and Bryan Lawrence³

1. Centre for Environmental Data Analysis, Science and Technology Facilities Council, Rutherford Appleton Laboratory
2. Scientific Computing Department, Science and Technology Facilities Council, Rutherford Appleton Laboratory
3. National Centre for Atmospheric Science, University of Reading

ABSTRACT

JASMIN is a highly successful data analysis system, which is used by thousands of academics and their industrial partners to analyse many petabytes of environmental data. The rapidly increasing volume of data stored on JASMIN, and the steadily increasing number of users, is making it necessary to investigate and implement new methods of providing computing resources to the users, storing the data that they produce from their analyses and storing and maintaining a very large archive of environmental data. To achieve this, two main areas of research are described. Firstly, providing users with virtualised services to best utilise the computing resources available. Secondly, using object storage to provide a large, yet affordable, data store and providing the users with tools and interfaces to common environmental data formats, so as to not unduly affect their current work flows.

1. INTRODUCTION

We describe key avenues of development for the next evolution of JASMIN [1], a hosted processing and data analysis facility for the UK environmental sciences community and its work with international partners. Now in its sixth year of operation, JASMIN provides a large curated archive of earth observation and climate science datasets, group workspaces for users to store their data, together with a batch compute environment (LOTUS) and a community cloud; all hosted on an infrastructure customised for high bandwidth and low latency between compute and storage.

2. DEVELOPMENT OF JASMIN

There are three main development areas: supporting the ongoing migration of compute workloads to containers, developing a cluster-as-a-service provision, and delivering additional tiers of storage. These all arise from the primary purpose of JASMIN: to provide a “data commons” where users can see existing data, add their own, and exploit their

own computational environments to manipulate the data. Usage of JASMIN is growing along three independent axes: number of users, volume of data stored, and number of communities supported. Throwing more hardware at the problem alone (whether in-house at JASMIN, or in the public cloud) cannot be afforded, and more sophisticated approaches are necessary, particularly to support less computationally mature user communities.

3. USER ENVIRONMENTS

One key challenge is the ability to make most effective use of parallelism in order to best utilise the computational resources available. This is impacted by usability: for some users, the technical expertise needed to using traditional batch compute is too high or they don't know how to refactor their code to make it work in parallel. Additionally, there is a need for interactive and graphical application environments in order to analyse data. Technologies such as Jupyter [2] and Zeppelin [3] Notebooks and Dask [4] and Apache Spark [5], when used in combination, provide a new opportunity to provide interactive data analysis and a means to exploit parallelism which to some degree abstracts the complexities from the user.

In order to deliver these environments for our user communities, we need a means to rapidly deploy the underlying virtualised infrastructure a so-called *cluster-as-a-service*. Our experiences with earlier projects on JASMIN such as the ESA-funded OPTIRAD [6] project have shown the potential for container technologies used in conjunction with such services. Recent work with Kubernetes [7] has demonstrated the ability to rapidly deploy environments, make services elastic, more effectively manage the allocation of resources between application and traditional cloud virtualisation tiers and also more easily port between different platforms. Nevertheless, this and other application scenarios must address challenges around storage: interfaces to it, performance and scaling.

4. DATA STORAGE

Most of the current JASMIN storage is delivered using a parallel file system. This gives both users and administrators an easily accessible, highly performant storage environment. However, data volumes are increasing rapidly, arising from both new satellite missions, and more environmental simulation, as well as an even more rapid increase in user-created data. With growing data volume comes the increasing need to make better use of tiered (and cheaper) storage for less frequently accessed data and/or data for which high-performance access is not required. At the same time, the new modes of interaction (from containers, and both private and public cloud) require new methods of interacting with the data. Recent work exploring the use of object storage, alongside new tape caching systems, has shown the potential to address many of these issues. In particular, object stores with simple HTTP REST interfaces such as Amazon S3 [8], provide an approach which avoids the limitations of mounting file systems and managing root privileges, yet can support higher level semantics and rich metadata with reasonable performance.

Even so, S3 presents a fundamental change in how users access data and presents challenges in how best to provide efficient access to data stored using binary formats such as the HDF5/NetCDF4 [9] data model. Work underway with the EU ESiWACE [10] project and collaboration with the HDFGroup [11] in the US has shown how it is possible to make an efficient system to store and access such data by distributing the storage of files across multiple objects with each object representing a fragment of data from the file. In such a way, it is possible to create a subsetting interface to extract individual content without downloading the whole of an object to a client. Using the HDF REST API [12] or the OPeNDAP [13] specification it is possible to integrate this interface with the standard NetCDF client libraries so that the interface presented to the user is largely unchanged.

In addition to HDF5/NetCDF4 files, JASMIN holds a large quantity of both structured and unstructured data, with a large volume of zip files in particular. Additional work will concentrate on how to exploit the internal structure of the structured files to enable them to be split across multiple objects, and finally to provide a solution for the storage of unstructured data. In all cases, deriving and holding rich metadata for each object is a primary goal, to enable searching both the data archive and user data without fetching whole objects.

5. ROADMAP TO IMPLEMENTATION

In order to implement an object storage based solution, a hierarchy of interfaces will be offered to users, with the most simple being implemented first:

1. Object storage used in conjunction with POSIX file system cache i.e. whole files stored in object store and

retrieved and operated on by applications from the cache - analogous to tape retrieval. A client utility will be provided to the users

2. Implement a user focused variant of the netCDF4 Python library enabling the splitting a netCDF file into smaller netCDF files using the Climate Format Aggregator (CFA) [14] conventions.
3. Implement a server focused solution to serving netCDF files from the archive, using the HDFGroup Highly Scalable Data Service (HSDS) [15].
4. Provide object storage support for other file formats, determined on which is most prevalent in the archive and user data.

6. REFERENCES

- [1] <http://www.jasmin.ac.uk>
- [2] <http://jupyter.org>
- [3] <https://zeppelin.apache.org>
- [4] <https://dask.pydata.org>
- [5] <https://spark.apache.org>
- [6] http://www.esa.int/Our_Activities/Space_Engineering_Technology/Shaping_the_Future/Optical_Multisensor_Radiance_Data_Fusion_Techniques_OPTIRAD
- [7] <https://kubernetes.io>
- [8] <http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html>
- [9] <http://www.unidata.ucar.edu/software/netcdf/docs/index.html>
- [10] <https://www.esiwace.eu/>
- [11] <https://www.hdfgroup.org/>
- [12] <https://support.hdfgroup.org/projects/hdfserver/-rest>
- [13] <https://www.opendap.org/>
- [14] <http://www.met.reading.ac.uk/~david/cfa/0.4/index.html>
- [15] <https://support.hdfgroup.org/projects/hdfserver/>

SEMANTIC SEGMENTATION USING DEEP NEURAL NETWORKS FOR SAR AND OPTICAL IMAGE PAIRS

*Wei Yao¹, *Dimitrios Marmanis^{1,2}, Mihai Datcu¹

¹Department of Photogrammetry & Image Analysis, IMF, German Aerospace Center (DLR), Germany
²Department Photogrammetry & Remote Sensing, Technische Universitaet Muenchen (TUM), Germany

ABSTRACT

Semantic segmentation for synthetic aperture radar (SAR) imagery is a rarely touched area, due to the specific image characteristics of SAR images. In this research, we propose a dataset which consists of three data sources: TerraSAR-X images, Google Earth images and OpenStreetMap data, with the purpose of performing SAR and optical image semantic segmentation. By using fully convolutional networks and deep residual networks with pre-trained weights, we investigate the accuracy and mean IOU values of semantic segmentation for both SAR and optical image patches. The best segmentation accuracy results for SAR and optical data are around 60% and 82%. Moreover, we study SAR models by combining multiple data sources: Google Earth images and OpenStreetMap data.

Index Terms— Deep learning, Semantic segmentation, TerraSAR-X, Google Earth, OpenStreetMap

1. INTRODUCTION

In remote sensing area, semantic segmentation has been always a challenging task, meanwhile, it's also a critical step for various applications. There are already a few researches focus on extracting different object from optical satellite-borne or air-borne data [1]; however, very rare cases have been studied for SAR images, or only for very simple applications [2].

In computer science area, the booming development of deep learning methods have shown great power in image information mining from big dataset. Currently, most for computer vision applications. For our remote sensing applications, it's reasonable to assume that, deep learning methods can be one of the ever powerful algorithms that have the potential to beat the other on-the-shelf algorithms (SVM, random forest, etc.).

Hence in this research, we investigate the potential of deep learning methods for a large amount of data, and present here our preliminary semantic segmentation results for high resolution SAR and optical images, based on deep learning methods. Specifically, our dataset contains 6000 image patches with a size of 200x200 pixels, which are labeled by four cat-

egories: building, natural, landuse and water. We are particularly interested in the extraction of buildings. We plan to experiment on different deep learning models, and modify the networks architecture as well, in order to see their effects on getting more accurate segmentation results. Moreover, with multiple sources of knowledge, we study the characteristics of SAR models.

We choose the well-known fully convolutional networks (FCN), together with 50 layer deep residual networks and Atrous convolution networks for our experiments.

In conclusion, the following points show our main contributions:

- We introduce a heterogeneous dataset which consists of TerraSAR-X imagery, Google Earth optical imagery, OpenStreetMap data for pixel-wise semantic segmentation with four categories.
- We build optical models and SAR models, based on deep learning Fully Convolutional Networks (FCN) and Deep Residual Networks learning scheme.
- We study SAR models (i.e., feature maps) by combining multiple data sources: Google Earth images and OpenStreetMap data.

2. METHODOLOGY

2.1. Dataset Description

Our Dataset consists of three data sources, which are TerraSAR-X GEC products, OpenStreetMap data, optical Google Earth images, with the same resolution of 2.9 meters. It covers 15 cities of North Rhine-Westphalia (NRW), Germany. The TerraSAR-X GEC products are with a ground resolution of 2.9 meters, and the incident angles are between 20 and 45 degree on various shooting dates and orbits. As the geocoded coordinates are provided by GEC products, we use open source data from OpenStreetMap to build the corresponding ground truth. This largely reduces the human labeling effort, however, there is also shortcomings, that quite an amount of pixels are without specific map information, due to the lack of geographical information from the open source data. Thanks to the error tolerant ability of deep learning methods, we can

*Authors have contributed equally in this work.

still use the majority data of our dataset. Besides, the corresponding optical data were downloaded from Google Earth. All three data sources are processed to the same resolution. Figure 1 shows an example of image patches from three heterogeneous data sources. For OpenStreetMap patch, black color stands for buildings, blue color stands for landuse, red color stands for natural, green color stands for water which is not shown in this example.



Fig. 1. Example of image patches from heterogeneous data sources.

2.2. Fully Convolutional Networks & Atrous Convolution Networks & Deep Residual Networks

Fully convolutional networks (FCN) are a kind of deep-learning neural networks which change the last fully connected layers of classification networks to fully convolutional networks. In such context, the networks are adjusted to solve a "pixel in, pixel out", namely segmentation problem [3]. They have shown great success in a number of computer vision and aerial remote sensing applications.

Atrous convolution, also known as dilated convolution, is a shorthand for convolution with upsampled filters. This idea have been used before in the context of DCNNs. In practice, atrous convolution computes feature maps more densely, in order to recover full resolution feature maps. Compared to regular convolution, atrous convolution allows us to effectively enlarge the view field of filters without increasing the number of parameters or the amount of computation [6].

Deep Residual Networks are a kind of very deep neural networks with many layers that have obtained impressive results recently. They have an intriguing "connection skipping" mechanism which enables the inputs of a lower layer available to a node in a higher layer [4].

2.2.1. Networks Explanation

For our experiments, we use the "voc-fcn8s", "FCN-ResNet50-16s" and "Atrous-ResNet50-16s" pre-trained models, specifically, the deep residual network is with 50 layers and a stride of 16 pixels for optical and SAR models. Caffe framework [5] and Keras which is based on Tensorflow are used to implement our deep-learning algorithms. By experimenting with different deep learning models, we will analyze their results from the quantitative and visualizing perspectives in the next chapter.

3. RESULTS AND ANALYSIS

In this section, results for the fully connected neural networks, deep residual neural networks are presented. Atrous residual networks have presented here.

3.1. Quantitative Results

For our experiments, we have used two quantitative results to evaluate our models: accuracy and mean IOU. The mean Intersection-Over-Union (mean IOU) is a common evaluation metric for image semantic segmentation. It computes the IOU for each semantic class, then computes the average over classes. The IOU is defined as:

$$\text{IOU} = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}} \quad (1)$$

Then a confusion matrix is obtained based on the predictions, and mean IOU is calculated from it.

Table 1. Segmentation accuracies for different data sources and models.

Model \ Data	TerraSAR-X 16bit	Google Earth
Atrous-ResNet50-16s	0.609	0.829
FCN-ResNet50-32s	0.461	0.827

Table 1 describes the segmentation accuracy values by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. Generally, Google Earth image patches get higher accuracies, with around 82% correct segmentation. It's interesting to notice that, regarding different models, there is a big difference for TerraSAR-X image patches, while almost no change for Google Earth image patches. This means the Atrous convolution matters a lot to TerraSAR-X data.

Table 2. Segmentation mean IOUs for different data sources and models.

Model \ Data	TerraSAR-X 16bit	Google Earth
Atrous-ResNet50-16s	0.269	0.437
FCN-ResNet50-32s	0.225	0.422

Table 2 describes the segmentation mean IOU values by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. The mean IOU values are much lower than segmentation accuracy results. Like accuracy, Google Earth image patches get better values. But regarding each data source, there is no significant difference between models.

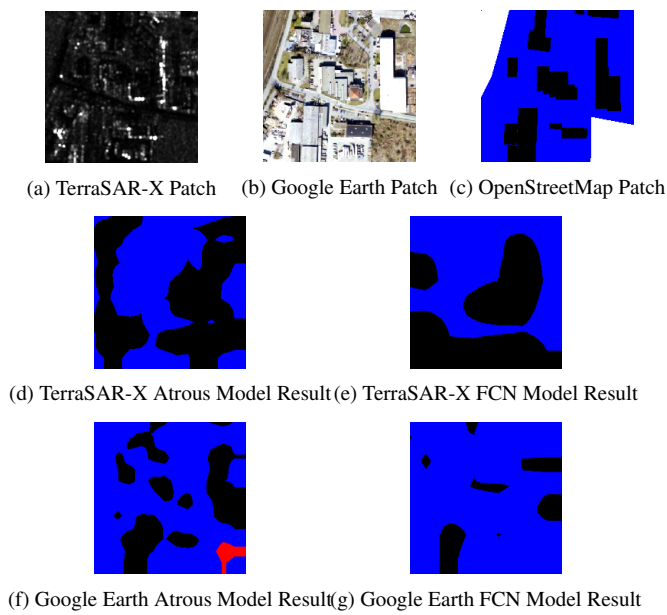


Fig. 2. Building example of Google Earth Patch Result and TerraSAR-X Patch Result.

3.2. Visualized Analysis

Figure 2 mainly shows the building segmentation results by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. TerraSAR-X results get the strong scattering locations which show as bright spots areas in the image patches. However, Google Earth result obtained from Atrous model detects building more precisely. Generally, results from Atrous model are better than from FCN model.

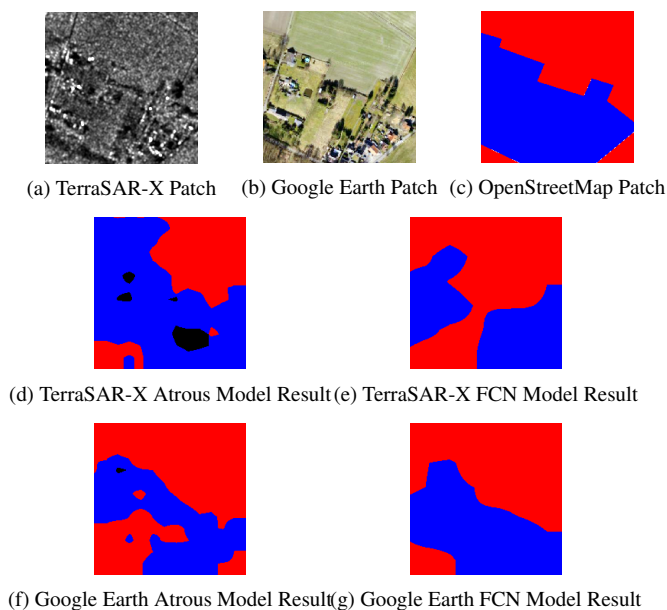


Fig. 3. Landuse example of Google Earth Patch Result and TerraSAR-X Patch Result.

Figure 3 mainly shows the landuse segmentation results by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. In case of landuse class, SAR results look better than optical results. This maybe due to the scattered strong scattering spots within the landuse area. Figure 3(e) and (f) actually correctly detect the center part of landuse is actually natural class, which indicates a combination of both SAR and optical data could bring a better result. Also due to the small-scale of buildings, comparing to Figure 2, they are difficult to be correctly segmented.

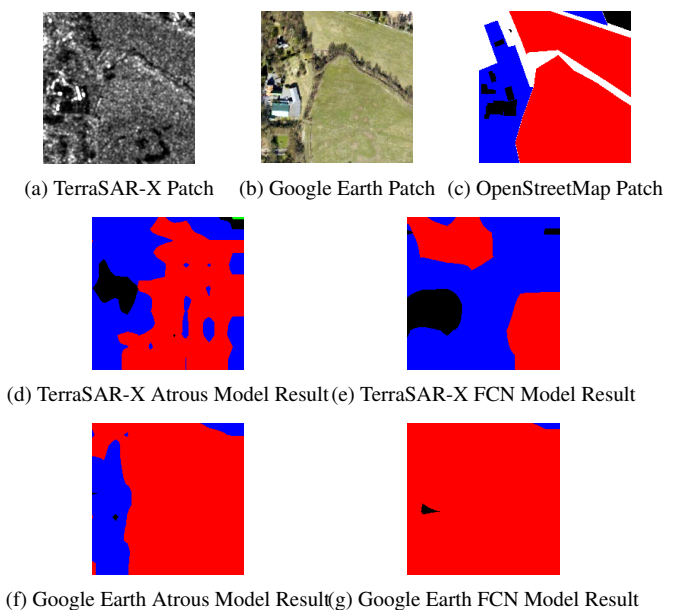


Fig. 4. Natural example of Google Earth Patch Result and TerraSAR-X Patch Result.

Figure 4 mainly shows the natural segmentation results by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. The natural class is relatively easy to be segmented, as the strong texture features it shows. For this example. Atrous model gets much better results than FCN model. For those middle-scale buildings, because of the strong scattering effect which inherited from SAR imaging mechanism, it's nice to see that SAR results get better building segmentation than optical results.

By comparing the building extraction results from Figure 2, Figure 3 and Figure 4, we found SAR model is relatively more sensitive to detect building areas.

Figure 5 mainly shows the water segmentation results by using the pre-trained Atrous residual network and FCN residual network for SAR and optical image patches. For water class, optical results show almost perfect segmentation; meanwhile SAR results get nothing due to the low image patch intensity value.

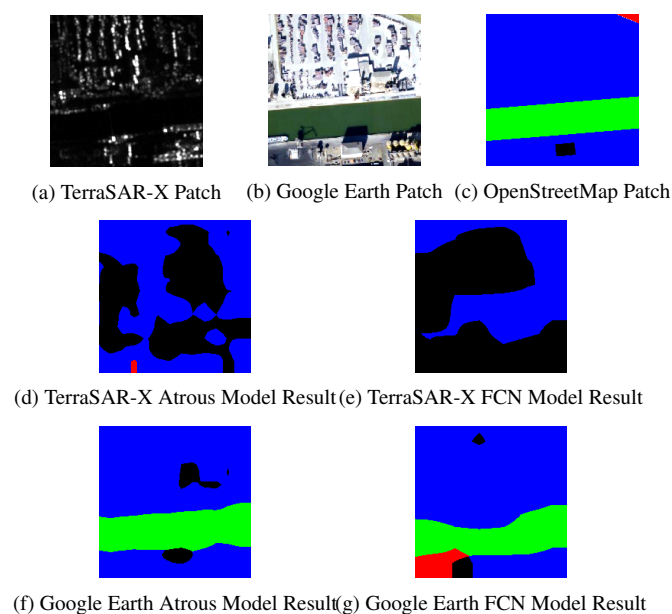


Fig. 5. Water example of Google Earth Patch Result and TerraSAR-X Patch Result.

3.3. Conclusions and Outlook

Comparing the state of the art semantic segmentation of earth observation data [7], [8], the results we obtained are not yet good enough, however, our contribution lies on analyzing both SAR and optical models and segmentation results simultaneously. Regarding the quantitative results, there is still a large space to improve. These are supposed to achieve by changing networks architecture and fine-tuning model parameters. Moreover, at the moment our results are separate for SAR and optical data. Inspired by the visualized analysis, it would be very interesting to see the results of a combination of both data sources.

Hence, here are our outlook to the future work:

Since we are facing a big data scenario, the segmentation results will be benefited by increasing the size of our dataset. Currently, we still have a number of the same type of high resolution GEC TerraSAR-X products and their corresponding Google Earth optical images, we plan to download the corresponding OpenStreetMap data to increase the volume of our dataset. In such context, more detailed categories could be considered, for example, natural category can be split up into forest, grass, farm, etc. And we study the potential of our networks models to extract more detailed information.

Furthermore, we will investigate the behaviors of layers (i.e., add layers, skip layers, connect to lower layer, etc.), and their impacts on increasing the semantic segmentation accuracy for our dataset. Then we will adjust the networks architecture to combine multiple sources of knowledge, with the purpose of obtaining better SAR models by training optical and map information together.

4. REFERENCES

- [1] D. Marmanis, J.D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNNs," in *ISPRS Annals of the Photogrammetry, remote sensing and Spatial Information Sciences*, Prague, Czech Republic, July 2016, vol. III-3.
- [2] W. Yao, S.Y. Cui, H. Nies, and O. Loffeld, "Classification of land cover types in TerraSAR-X images using Copula and speckle statistics," in *Proceedings of the 10th European conference on Synthetic Aperture Radar (EUSAR 2014)*, 2014, pp. 743–746.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transaction of Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, pp. 640–651, 2017.
- [4] K.M. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Deep residual learning for image recognition," *Computing Research Repository (CoRR)*, vol. abs/1512.03385, 2015.
- [5] Y.Q. Jia, E. Shelhamer, J. Donahue, and S. Karayev, "Caffe: Convolutional architecture for fast feature embedding," *ACM multimedia 2014 open source software competition*, 2014.
- [6] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Re-thinking atrous convolution for semantic image segmentation," *Computing research repository (CoRR)*, vol. abs/1706.05587, 2017.
- [7] A. Lagrange, Beaupere A. Le Saux, B., A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: from learning explicit features to convolutional networks," in *IEEE International Geosciences and Remote Sensing Symposium (IGARSS)*, 2015, pp. 4173–4176.
- [8] N. Audebert, Le Saux B., and Lefèvre S., "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," *Computing research repository (CoRR)*, vol. abs/1609.06846, 2016.

ARTIFICIAL GENERATION OF BIG DATA FOR IMPROVING IMAGE CLASSIFICATION: A GENERATIVE ADVERSARIAL NETWORK APPROACH ON SAR DATA

*Dimitrios Marmanis^{1,3}, *Wei Yao¹, Fathalrahman Adam¹, Mihai Datcu¹,
Peter Reinartz¹, Konrad Schindler², Jan Dirk Wegner², Uwe Stilla³

¹Department of Photogrammetry & Image Analysis, German Aerospace Center (DLR), Germany

²Photogrammetry & Remote Sensing Group, ETH Zurich, Switzerland

³Department Photogrammetry & Remote Sensing, Technische Universitaet Muenchen (TUM), Germany

ABSTRACT

Very High Spatial Resolution (VHSR) large-scale SAR image databases are still an unresolved issue in the Remote Sensing field. In this work, we propose such a dataset and use it to explore patch-based classification in urban and peri-urban areas, considering 7 distinct semantic classes. In this context, we investigate the accuracy of large CNN classification models and pre-trained networks for SAR imaging systems. Furthermore, we propose a Generative Adversarial Network (GAN) for SAR image generation and test, whether the synthetic data can actually improve classification accuracy.

Index Terms— Big Data, SAR classification, GANs, Generative Adversarial Networks, Deep Learning

1. INTRODUCTION

Classification of very high resolution (VHR) SAR image data remains a hard and time-consuming task. Major difficulties include the scarcity of available data, and the challenge of semantically interpreting the SAR backscatter signal. Linked to those difficulties, there are no large-scale, SAR-derived image databases for Remote Sensing image analysis and knowledge discovery. Furthermore, while optical image classification has seen a breakthrough with the advent of *Deep Learning* methods that require Big Data, SAR-based systems have so far not experienced the same progress, likely because of not enough data with associated training labels is available.

In this work we try to tackle the lack of training data, by introducing a large-scale SAR image database. Precisely, our dataset contains more than 60'000 image instances and respective labels, chosen from 7 distinct semantic classes. Using this data, we perform a set of experiments to understand the impact of dataset size on classification accuracy. In this context, we also investigate the possibility to further expand the dataset with synthetic SAR images generated with the help of *Generative Adversarial Networks (GANs)*. These are powerful generative models that have been shown to produce

high-quality synthetic images in other fields, thereby reducing (or even completely avoiding) the annotation effort. Our main contributions in this work can be summarized as follow:

- We construct the first state-of-the-art CNN model pre-trained on large-scale SAR data.
- We investigate the possibility of transfer-learning from other pre-trained models based on optical images, and their impact on SAR image classification.
- We investigate the possibility of training also with artificial SAR data generated with a GAN.

2. RELATED WORK

In the field of SAR image analysis, the use of deep-learning methods, such as *CNNs*, is still in its infancy, mainly due to the limited availability of VHR data with associated ground truth labels. We note that, in a detailed literature review, we did not find any work that relies on a large scale SAR-database to unlock the potential of deep neural networks. Moreover, there are no pre-trained networks for SAR images, which would facilitate the classification of SAR datasets for which there aren't enough training labels to learn a deep network from scratch.

Published work at the intersection of SAR imaging and deep learning are mainly focussed on *Target Classification*. Some representative works employ sparsely connected layers [1], limited training data [2] and domain-specific data augmentation methods [3]. In the field of *GANs* for SAR data, some interesting results have been shown by [4], where authors constructed a generative deep model. The outcome of their experiments however remain inconclusive, due to the scarcity of training data, and particular characteristics of the underlying targets (military imagery). Another implementation of *GANs* in the field of Remote Sensing is the one of [5], who investigate the *Wasserstein GAN* for poverty mapping with sparse labels, using a semi-supervised approach. They however do not use SAR imagery. Yet another work

*Authors have contributed equally in this work

on optical remote sensing imagery and artificial data generation is the one of [6]. They propose an additional objective function over the standard GAN architecture to improve the output. While the approach is interesting, it ultimately does not produce visually realistic images of the target classes. A promising work is [7], which demonstrates the generation of synthetic SAR images on the basis of optical images. The high-quality samples generated in that work show the potential of GAN methods for SAR image synthesis, and motivate us to further investigate that topic.

3. THE DATASET

Our dataset was obtained via a novel classification scheme especially designed for high-resolution SAR imagery of (mainly) built-up areas. The dataset contains image patches from 288 TerraSAR-X image scenes (41 scenes acquired in Africa, 6 from Antarctica, 59 from Asia, 80 from Europe, 40 from the Middle East, 54 from North and South America and 8 from ocean surfaces), with a total of over 60'000 individual patches. All TerraSAR-X data are obtained via the X-band instrument, using the high-resolution Spotlight mode. The incident angles throughout the scenes varies between 20 and 50 degrees. The resolution of the images scenes is set to 2.9m, with a pixel spacing of 1.25m. The chosen polarization for the dataset is horizontal (HH) for all products. Furthermore, for convenience we convert all intensity data to 8-bit integer precision. For more information on the dataset, refer to [8].

4. EXPERIMENTS

In our experiments, we first set a baseline for deep learning based SAR classification, and go on to investigate if we can improve over that baseline with additional, synthetic data generated with a GAN.

4.1. The CNN SAR classifier

To establish a baseline for the use of CNNs with SAR data, we employ a state-of-the-art network architecture for optical images, namely the standard Residual Network with 50 hidden layers (*ResNet-50*) [9]. To adapt the network to our class nomenclature, we remove the fully connected layers at the top and replace them with three fully connected layers of size 256, 256 and 7, respectively, which we train from scratch. The resulting model achieves an overall accuracy of 93.2%. We find this result very encouraging: in spite of the radically different imaging process and image statistics, modern, deep CNNs appear to be suitable for supervised SAR image classification and yield high classification accuracy, when trained on an appropriate, large training set.

A further, interesting observation is that conventional pre-training (i.e., initialization with the weights learned from optical images) has little effect on the classification result. This

is not unexpected – while the pre-training with very large databases (millions of images) does usually help when working with optical images, the local image statistics of RGB and SAR data are probably too different to transfer even low-level image properties. To support that hypothesis, we have trained the same ResNet-50 twice, once with random initialization and once with weights pre-trained on *ImageNet*. The classification results for SAR were practically the same in both cases. I.e., the pre-trained weights do not hurt the learning, but they also do not help compared to random initialisation.

4.2. Image Generation with BEGAN Models

Given the good performance of the deep network, and the still comparatively small training database (in computer vision, models are routinely pre-trained with more than 10^6 training images), we investigate if artificial data generation with a GAN can further improve our classifier. For close-range applications, it has already been shown that classifier training can benefit from GAN image synthesis, e.g., for sign recognition [10]. However, our task however is more challenging, due to the extreme variability of the SAR data in our database, and the large dimension of the output images we need to generate (160×160 pixels).

4.2.1. BEGAN Model for SAR

Despite the rather recent invention of GANs, there is already a plethora of variants such as *DC-GANs*, *cGANs*, *WGANs*, *DRAGANs* and *BEGANs*. We base our investigation on the newly proposed *BEGAN* model [11], which was shown to generate images of remarkable quality, and to handle larger image sizes than most other variants.

Compared to the standard GAN model, the *BEGAN* design has a number of attractive characteristics. First, it uses autoencoders as discriminator, thus matching the corresponding autoencoder distributions (rather than the raw data distributions), with a Wasserstein distance loss. Furthermore, *BEGAN* employs an equilibrium term to balance the effect of the *Discriminator* with respect to the *Generator*, so as to avoid an “early win” of one stage over the other.

BEGAN was initially proposed for generating human faces. Even though this is already a challenging problem, synthesising SAR images proved to be a lot harder. Through empirical experimentation, we found that the capacity of the original model is not sufficient to capture the complexity of our database. We therefore added more layers both to the *Generator* and the *Discriminator*. In each of the two stages, we add two additional convolution layers (with respective eLU non-linearities), before the respective pooling/upsampling layers. Furthermore, we have replaced the final, linear layers of both stages with non-linear ones, using the ReLU non-linearity.¹

¹Code: https://github.com/deep-unlearn/Big_Data_From_Space_2017

Finally, and perhaps most significantly, we have changed the loss function of the discriminator. The original loss function is simply the mean of the per-pixel L_1 distance. In our model, we replace it by a combination of a per-pixel distance and a histogram distance, to explicitly match the global intensity distributions of the images. The new loss is given by :

$$\begin{aligned}\mathcal{L}_{\text{generated}} &= L_{\text{hist}} + \omega \cdot L_{\text{spatial}} \\ L_{\text{hist}} &= \frac{1}{N_{\text{bins}}} \cdot \sum (\text{hist}(X) - \text{hist}(X_{\text{recon}}))^2 \\ L_{\text{spatial}} &= \frac{1}{N_{\text{pix}}} \cdot \sum (X - X_{\text{recon}})^2,\end{aligned}$$

where *hist* returns the histogram of an image over a fixed number N_{bins} of bins (set to 64), and N_{pix} is the number of pixels in the generated image. The hyperparameter ω defines a weighting between the two parts of the loss. For our experiments we empirically set it to $\omega = 0.001$.

4.2.2. BEGAN Image Generation

Image generation with GANs still remains somewhat a brittle and somewhat challenging task. We thus investigate three for our SAR image generation problem. They are:

- In the hard scenario, the network is asked to directly generate large SAR patches of size 160×160 pixels. This scenario would be optimal, in the sense that it outputs patches at the correct size for our database; but it also the most complex prediction task.
- In the intermediate scenario, the network generates SAR patches at $2 \times$ larger GSD, with dimension 80×80 pixels, which are then compared to downsampled real images. The reduced resolution lowers the complexity of the task, while the patch size in scene coordinates, and thus the spatial context, remains the same. But the resulting images must be upsampled to the original dimensions, and thus lack high-frequency detail.
- In the simple scenario, images are also generated at 80×80 pixels, but this time the original GSD is retained. Instead, the patch size in scene coordinates is halved, respectively the real SAR patches are cropped. The resulting images must again be upsampled, to match the patch size used for classification. Using smaller and more local patches presumably further reduces the complexity of the prediction, the price to pay is a mismatch in GSD between synthetic and real training images, and the loss of $3/4$ of the context area.

So far, we were unsuccessful in our attempts to train the hard scenario. We leave it to future work to determine whether this can be remedied, or whether a higher-capacity model is needed. For the intermediate scenario, the generator appeared to converge better, but its outputs were still unsatisfactory and did not visually resemble the original data. For the time being, this failure leaves us with the simple scenario. That setting did

converge to a reasonable solution that outputs realistically-looking synthetic images, see examples in *Figure 1* and real SAR data in *Figure 2*. However, one can also clearly see that the smaller patches capture less of the context.

4.3. Classification Augmentation Through GANs

In spite of the limited success to synthesize full-size patches, we continued the experiment. The “simple” patches were upsampled to 160×160 pixels and added to the training data for the classification network. As a first test, we generated 5100 synthetic instances of the *Settlement* class, which is the most frequent class in the dataset (25'000 real training patches), and also the one with the strongest intra-class variation.

Somewhat surprisingly, retraining the *ResNet-50* classifier with the augmented dataset did not influence the classifier either way. We get the same classification accuracy of 93.2%. Seemingly, the synthetic examples were neither capable of adding any additional information that would have improved the classifier, nor were they unrealistic enough to negatively impact the classifier. Obviously, in the absence of a satisfactory explanation such an outcome appears unlikely. Future work will have to determine the cause, and hopefully address the current short-comings of the generator, so as to further improve the classifier network.

5. CONCLUSIONS

We have introduced a new, large-scale database of SAR patches with associated semantic class labels. To our knowledge, this is the first SAR dataset large enough to train modern deep neural networks, and we have demonstrated that capability by learning a *ResNet-50* convolutional network that achieves an excellent 93.2% hit rate over 7 different scene categories. We have further adapted the generative BEGAN network model to SAR data, and have experimented with synthetically generated images to obtain an even larger training set. Unfortunately, we are still struggling with technical problems in the image synthesis, and the first experiments with additional, synthetic training data have not yet led to conclusive results. Nevertheless, our paper clearly shows that, as soon as enough data is available, deep convolutional networks work extremely well also for SAR images. More detailed tests and comparisons still need to be run, but we believe that our results set a new standard for patch-wise SAR classification. We also posit that our failure to exploit synthetic images is due to relatively minor technical difficulties that can be addressed, and we are still convinced that GANs have the potential to support the the generation of truly big training databases.

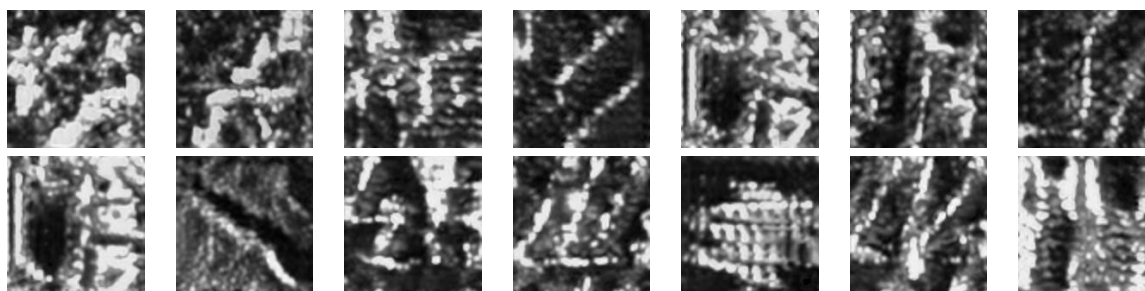


Figure 1. Generated data of size 80×80 pixel by cropping scenario - upsampled to 160×160 pixel

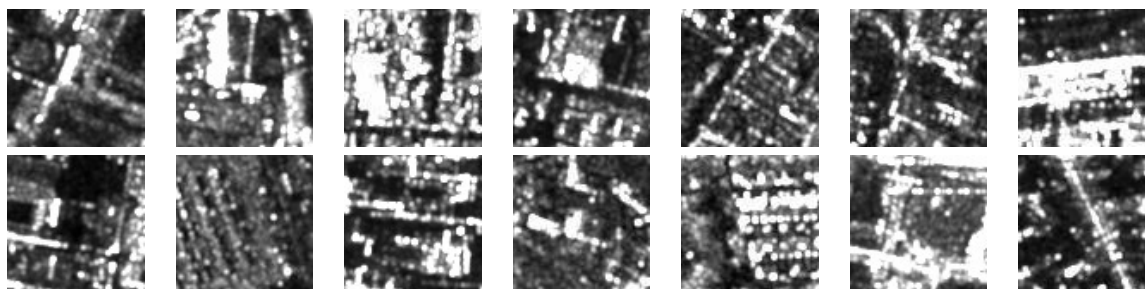


Figure 2. Original TerraSAR-X data of original size - 160×160 pixel

6. REFERENCES

- [1] Sizhe Chen, Haipeng Wang, Feng Xu, and Ya-Qiu Jin, "Target classification using the deep convolutional networks for sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4806–4817, 2016.
- [2] Zhao Lin, Kefeng Ji, Miao Kang, Xiangguang Leng, and Huanxin Zou, "Deep convolutional highway unit network for sar target classification with limited labeled training data," *IEEE Geoscience and Remote Sensing Letters*, 2017.
- [3] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang, "Convolutional neural network with data augmentation for sar target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 3, pp. 364–368, 2016.
- [4] Jiayi Guo, Bin Lei, Chibiao Ding, and Yueting Zhang, "Synthetic aperture radar image synthesis by using generative adversarial nets," *IEEE Geoscience and Remote Sensing Letters*, 2017.
- [5] Anthony Perez, Swetava Ganguli, Stefano Ermon, George Azzari, Marshall Burke, and David Lobell, "Semi-supervised multitask learning on multispectral satellite images using Wasserstein generative adversarial networks (Gans) for predicting poverty," *Technical Report*, 2017.
- [6] DaoYu Lin, "Deep unsupervised representation learning for remote sensing images," *arXiv preprint arXiv:1612.08879*, 2016.
- [7] Nina Merkle, Peter Fischer, Stefan Auer, and Rupert Müller, "On the possibility of conditional adversarial networks for sar template generation," in *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. IEEE, 2017.
- [8] Corneliu Octavian Dumitru, Gottfried Schwarz, and Mihai Datcu, "Land cover semantic annotation derived from high-resolution sar images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 6, pp. 2215–2232, 2016.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [10] Xinlong Wang, Mingyu You, and Chunhua Shen, "Adversarial generation of training examples for vehicle license plate recognition," *arXiv preprint arXiv:1707.03124*, 2017.
- [11] David Berthelot, Tom Schumm, and Luke Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

SEA LEVEL ANOMALY PREDICTION USING RECURRENT NEURAL NETWORKS

Anne Braakmann-Folgmann, Ribana Roscher, Susanne Wenzel, Bernd Uebbing and Jürgen Kusche

Institute of Geodesy and Geoinformation, University of Bonn, Nussallee 15, D-53115 Bonn.

Contact: {abraakmann / rroscher / susanne.wenzel / bernd.uebbing / kusche}@uni-bonn.de

ABSTRACT

Sea level change, one of the most dire impacts of anthropogenic global warming, will affect a large amount of the world's population. However, sea level change is not uniform in time and space, and the skill of conventional prediction methods is limited due to the ocean's internal variability on timescales from weeks to decades. Here we study the potential of neural network methods which have been used successfully in other applications, but rarely been applied for this task. We develop a combination of a convolutional neural network (CNN) and a recurrent neural network (RNN) to analyse both the spatial and the temporal evolution of sea level and to suggest an independent, accurate method to predict interannual sea level anomalies (SLA). We test our method for the northern and equatorial Pacific Ocean, using gridded altimeter-derived SLA data. We show that the used network designs outperform a simple regression and that adding a CNN improves the skill significantly. The predictions are stable over several years.

Index Terms— sea level, neural networks, CNN, RNN, deep learning, climate change, altimetry, time series analysis

1. INTRODUCTION

Modelling and predicting sea level anomalies (SLA) is currently a relevant topic, as sea level responds to global warming directly by thermal expansion of the ocean and indirectly by mass increase through melting ice sheets and glaciers. And with 60% of the world's population living in coastal areas an accurate prediction of SLAs at a high spatial resolution is required. Sea level rise is both spatially and temporally highly variable and therefore especially hard to model. Predictions by ocean models rely on incomplete representation of physical processes, limited spatial resolution, uncertain initial conditions, and scenario-based boundary conditions.

Complementing physical ocean modelling, we develop a new approach using neural networks. Neural networks are a powerful mean to solve tasks such as classification, object detection and speech recognition, where they show promising results reaching accuracies superior to classical and shallow state-of-art machine learning algorithms [10].

So far, they have barely been explored in the context of sea level prediction. Makarynskyy et al. [6] use a single tide gauge to predict the local sea level in a harbour, and Wenzel and Schröter [13] use global tide gauge data, to derive average trends and amplitudes for eight different regions. However, they lack a high spatial resolution and the used one-layered fully connected network architectures are simple compared to current networks applied in the computer vision community.

Here, in contrast mostly convolutional neural networks (CNNs) and very deep networks are used, as they reach higher accuracies [10]. CNNs are particularly suitable for gridded data exploiting the spatial correlations. On the other hand recurrent neural networks (RNNs) and their enhancement Long Short Term Memory networks (LSTM, [3]) are specifically designed to model time series data. A combination of CNN and RNN has for example been used to predict one frame ahead in video sequences [8]. With sequence-to-sequence LSTMs [12] it is possible to predict several time steps ahead.

Our aim is first to design a sophisticated network composed of a CNN and an RNN to capture both spatial and temporal relations of radar-altimetric SLA fields in the Pacific Ocean and to predict the SLA map of the next month. Then we extend the prediction length to several years.

2. DATA

We use time series of SLAs from ESA (European Space Agency) CCI (Climate Change Initiative) [7] as input to our network. SLAs are the difference between actual sea surface height (SSH) and mean sea surface height. We use the Level 4 product of ESACCI, where all altimetry mission measurements have been merged into monthly grids with a spatial resolution of 1/4 degree. The data cover 23 years from January 1993 to December 2015. To validate the predictive skill, we split the data into training- (16 years), validation- (4 years) and test data (3 years).

The northern and equatorial Pacific make a perfect test region, since parts of it undergo rapid sea level rise while being affected by strong interannual and decadal variability [9]. Here we examine the region between 110° and 250° longitude and 15°S to 60°N latitude. This region holds 170,240 grid points, which give us almost 33 million values for training of the network and 47 million values overall.

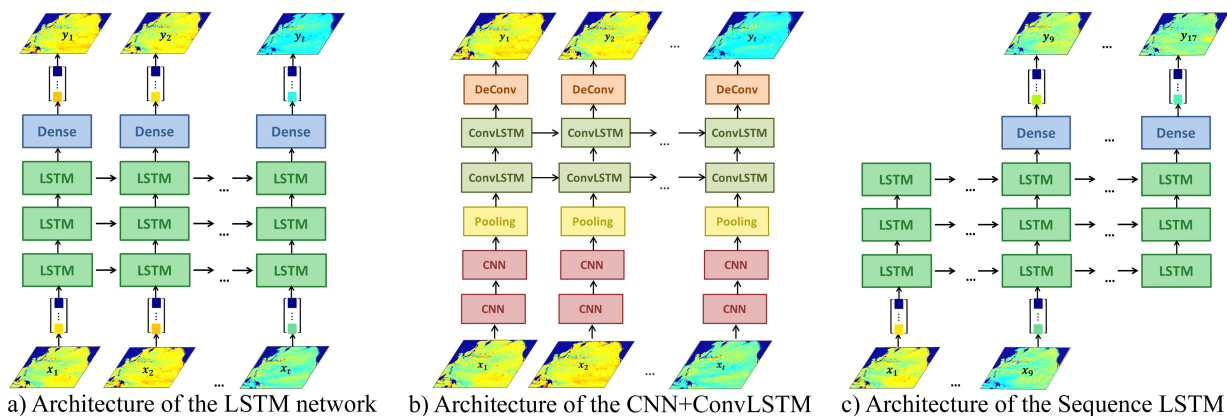


Fig. 1. Different network designs. The input x is all SLAs at one time step t . In a) and c) they are reshaped to a vector, in b) the grid structure is preserved. The network predicts the vector/grid y at the next one month (in a) and b)) or nine months (in c))

3. METHODOLOGY

SLAs reflect various complex non-linear interactions in the ocean and hence contain several deterministic and stochastic modes that are difficult to predict. Yet, SSH maps contain many reoccurring spatial patterns (e.g. eddy fields, gyres, ENSO) that can potentially be learned by neural networks.

Artificial neural networks are a means of machine learning inspired by the human brain to learn higher-order representations and perform diverse tasks. In contrast to other machine learning techniques, neural networks are able to extract relevant features and their weight in the model. CNNs have shown to be efficient networks which are especially designed to capture spatial dependencies by using gridded input data.

RNNs are neural networks that can deal with time series analysis, taking sequences as in- and output [12]. In contrast to feedforward networks they incorporate a self-loop, which enables the net to memorise the previous inputs (horizontal information flow in Fig. 1). LSTMs [3] are a special kind of RNN that are even capable of learning long-term dependencies. Convolutional LSTMs (ConvLSTM, [11]) preserve the input's spatial structure by replacing all matrix multiplications within the LSTM with convolutions. To learn higher-order representations, both RNNs and CNNs can be stacked on top of each other.

4. EXPERIMENTAL SETUP

We design different models using the framework Keras [1] with Tensorflow backend. We train them with the Adam optimizer [4] to minimise the MSE between their prediction and the truth in 150 epochs using decaying learning rates.

LSTM network: Here the input is a vector with all SLAs of one month. The network consists of three LSTM layers with 60 units each, followed by a fully connected layer that learns how to make a prediction of the next month for each $1/4^\circ$ grid cell from the 60 features (Fig. 1a). This prediction is

our output. The network consists of over 51 million parameters to train. We employ a hard sigmoid activation function for the hidden-to-hidden state translations and a tanh activation function in the LSTMs' hidden-to-output transformation, allowing the net to learn non-linear relations. To further improve the generalisation capability we employ LSTM-modified dropout [2] of 0.8. On an Intel i5-2400 CPU training takes 8 hours. After each prediction of one month we use the true values of the preceding month as input for the next prediction.

CNN+ConvLSTM network: Here the input SLAs are used in their natural grid structure, preserving spatial information. In this network we combine a CNN with a ConvLSTM as shown in Fig. 1b. The CNN consists of two convolutional layers with 32 filters, a kernel size of 3×3 and a ReLU activation function each, followed by a pooling layer with kernel size 4×4 and stride 4. We use the extracted feature maps as input to two ConvLSTM layers with 40 units each, a kernel size of 3×3 and the same activation functions as used in the LSTM. The last layer is a deconvolutional (DeConv, [14]) layer to regain the spatial extent of our input and to map the 40 feature maps to a single SLA value at each grid cell. This map of SLAs is our output prediction for the next month. Each predicted grid cell has a receptive field of 6° . This net has to learn only 229,413 parameters due to pooling. We apply batch normalisation to the input. Dropout is not used. On an Intel i5-2400 CPU training this net takes 18 hours.

Sequence LSTM network: This network is equal to the LSTM network, but takes sequences of nine months as in- and output (see Fig. 1c). While the former two networks only predict one month in advance, this network is able to predict the following nine months at once. After the prediction of a nine months long sequence, we use the true values of the preceding nine month as input for the next prediction. On the Intel i5-2400 CPU training this net takes 8 hours, too.

Sequence LSTM-P network: This network is trained just like the Sequence LSTM. However here we take the predicted nine months as input to predict nine months further.

5. RESULTS

In this section we assess the networks' performance and compare the predictions made by our networks to a regression approach. We use the sum of a trend, acceleration, annual and semi-annual sine and cosine as regression model. The regression parameters are estimated for each grid cell individually.

Figure 2 shows the measured time series of SLAs at one point in the North Pacific Gyre. The position is marked with a cross in Fig. 3. We visualize the prediction during training (blue), validation phase (green) and the actual test phase (red), compared to the known true data (black).

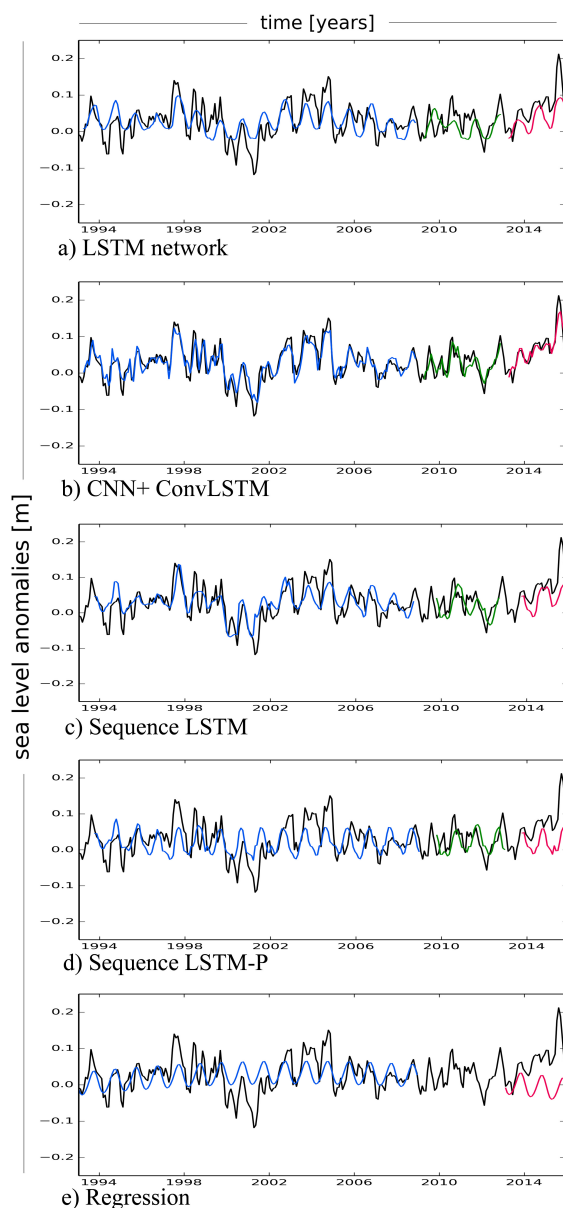


Fig. 2. Results for one grid cell (marked by a cross in Fig. 3): True (black) and predicted (blue = training, green = validation and red = test data) time series of SLAs

RMSE averaged over	training data	validation data	test data
LSTM network	0.062 m	0.071 m	0.076 m
CNN+ConvLSTM	0.047 m	0.050 m	0.051 m
Sequence LSTM	0.059 m	0.079 m	0.077 m
Sequence LSTM-P	0.083 m	0.080 m	0.081 m
Regression	0.078 m	-	0.154 m

Table 1. RMSEs between the networks' or the regression's prediction and the true SLAs averaged over all grid cells

Table 1 shows the RMSEs between true SLAs and the predictions made by our networks or the regression. The training error of the Sequence LSTM-P is higher, because we here feed the predictions 20 times back into the network.

To examine the spatial patterns, we plot the true and predicted SLAs in November 2014 in Fig. 3. All networks are trained using data up to 2008. Starting the test phase in January 2013, the Sequence LSTM-P needs nine true months as input only once (i.e. till September 2013). The Sequence LSTM depends on true data every nine months, so June 2014 is the last true month needed to predict the sequence including November 2014. The LSTM and CNN+ConvLSTM depend on true inputs of the previous month (here October 2014).

Striking, overall all network architectures outperform the regression. Especially in Fig. 3f it can be seen that the regression fails at capturing the spatial structure. We observe that an additional CNN improves the accuracy significantly (compare Fig. 2a/b and 3b/c). In Fig. 3c the CNN+ConvLSTM network resolves nearly all spatial structures very well. However, we observe that in combination with a normal LSTM, the CNN brings no improvement (not shown here) - only if we also keep the spatial structure throughout the ConvLSTM.

Both extending the prediction length to nine months (compare Fig. 2a/c and 3b/d) and using the predicted values for the next prediction (compare Fig. 2c/d and 3d/e) leads to only slight degradation of predictive skill and the predictions stay close to the real values for a long time.

Figure 4 shows the spatial distribution of the RMSE averaged over the test phase for the CNN+ConvLSTM and the Sequence LSTM-P. In both cases the main errors occur around the Kuroshio current with its highly variable eddies.

6. CONCLUSION

In this work we develop a combined convolutional and recurrent neural network to analyse both the spatial and the temporal evolution of SLAs in the northern and central Pacific Ocean and to make accurate predictions for the future sea level. We validate the accuracy of our approach and compare the results to a regression. All the used network designs outperform the regression. Adding a convolutional neural network improves the accuracy significantly. Therefore we plan to extend this architecture to take sequences as in- and output. We are confident that in this way we will be able to improve predictions at longer timescales. The developed method could also be applied to other regions or even globally.

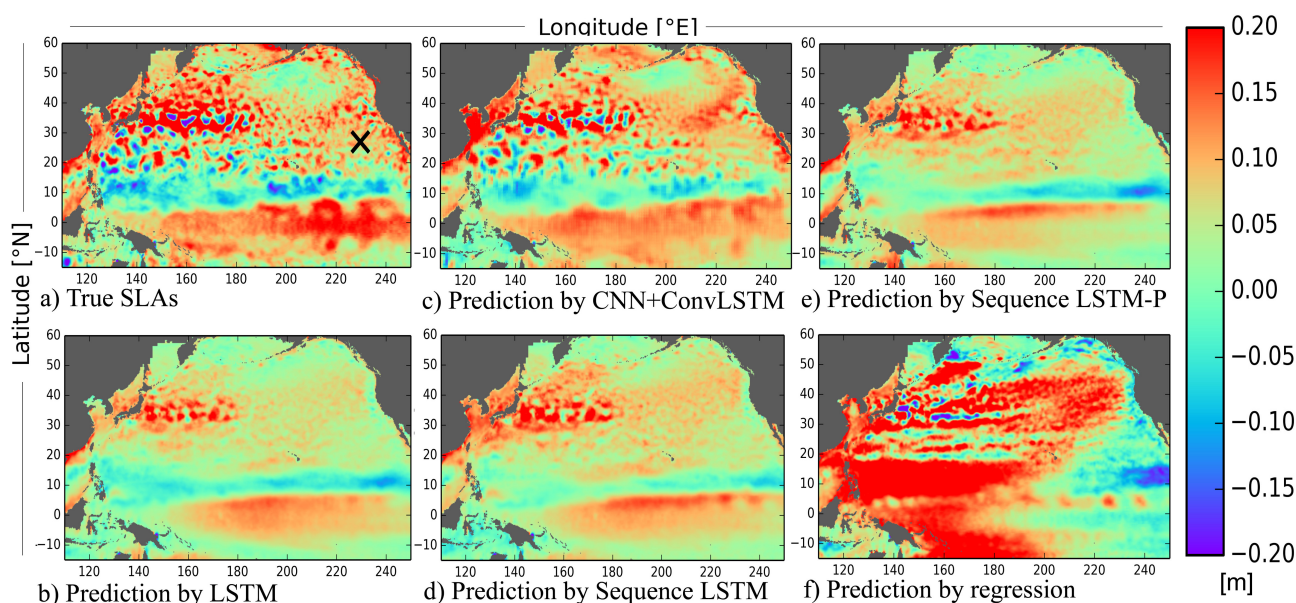


Fig. 3. Spatial structure of the SLAs in the selected region in November 2014 (within the test period)

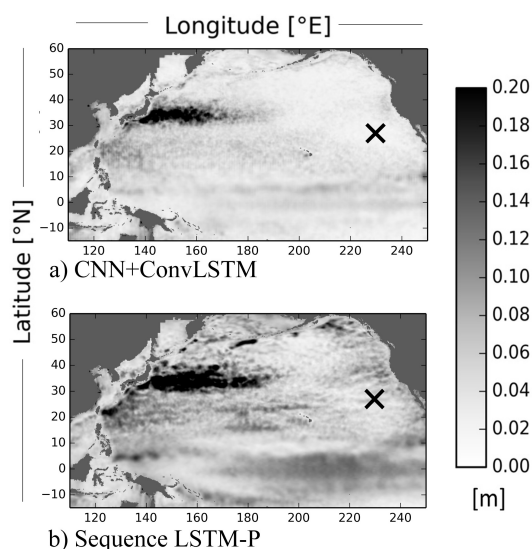


Fig. 4. RMSE at each grid cell averaged over the test phase

7. REFERENCES

- [1] F. Chollet et al.: Keras. Published on GitHub (<https://github.com/fchollet/keras>), (2015)
- [2] Y. Gal and Z. Ghahramani: A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. In NIPS, (2016)
- [3] S. Hochreiter and J. Schmidhuber: Long short-term memory. In Neural computation, (1997)
- [4] D. P. Kingma and J. Lei Ba: ADAM: A Method for Stochastic optimization. In ICLR, (2015)
- [5] P. Khorrami et al.: How Deep Neural Networks Can Improve Emotion Recognition on Video Data in Image Processing. In ICIP, (2016)
- [6] O. Makarynsky et al.: Predicting sea level variations with artificial neural networks at Hillarys Boat Harbour, Western Australia. In Estuar. Coast. Shelf Sci., (2004)
- [7] G. D. Quartly et al.: A new phase in the production of quality-controlled sea level data. In Earth Syst. Sci. Data Discuss., (2017)
- [8] M. Ranzato et al.: Video (language) modeling: a baseline for generative models of natural videos. In CVPR, (2015)
- [9] R. Rietbroek et al.: Revisiting the contemporary sea-level budget on global and regional scales. In PNAS, (2016)
- [10] J. Schmidhuber: Deep Learning in Neural Networks: An Overview. In Neural Networks, (2015)
- [11] X. Shi, Z. Chen, H. Wang and D.-Y. Yeung: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In NIPS, (2015)
- [12] I. Sutskever, O. Vinyals and Q. V. Le: Sequence to sequence learning with neural networks. In NIPS, (2014)
- [13] M. Wenzel and J. Schröter: Reconstruction of regional mean sea level anomalies from tide gauges using neural networks. In JGR, (2010)
- [14] M. D. Zeiler, D. Krishnan, G. W. Taylor and R. Fergus: Deconvolutional Networks. In CVPR, (2010)

DEEP SELF-TAUGHT LEARNING FOR REMOTE SENSING IMAGE CLASSIFICATION

Anika Bettge, Ribana Roscher, Susanne Wenzel

Remote Sensing, Institute of Geodesy and Geoinformation, University of Bonn, Germany, 53115 Bonn
s7anbett@uni-bonn.de, ribana.roscher@uni-bonn.de, wenzel@igg.uni-bonn.de

ABSTRACT

This paper addresses the land cover classification task for remote sensing images by deep self-taught learning. Our self-taught learning approach learns suitable feature representations of the input data using sparse representation and under-complete dictionary learning. We propose a deep learning framework which extracts representations in multiple layers and use the output of the deepest layer as input to a classification algorithm. We evaluate our approach using a multispectral Landsat 5 TM image of a study area in the North of Novo Progresso (South America) and the Zurich Summer Data Set provided by the University of Zurich. Experiments indicate that features learned by a deep self-taught learning framework can be used for classification and improve the results compared to classification results using the original feature representation.

Index Terms— self-taught learning, deep learning, archetypal analysis, landcover classification, remote sensing

1. INTRODUCTION

Classification of remote sensing images is an important task for land cover mapping. Recently, deep learning has become a valuable approach particularly for such classification tasks. As already pointed out by [1], the most successful approaches are supervised deep learning frameworks using a huge amount of labeled data for training. However, labeled data are scarce for remote sensing applications. In contrast, huge amounts of unlabeled data are available and easy to acquire.

In our approach we use self-taught learning (STL, [2]) which has turned out as a valuable procedure to the combined exploitation of unlabeled and labeled data, without the constraint that both datasets need to follow the same distribution. Therefore, we can utilize datasets from further scenes and acquisition times for feature/representation learning.

The most common approach to STL is sparse representation (SR), which learns features in an unsupervised way in order to use them for supervised classification. Some approaches use deep sparse representations DSR, which shows improved results over shallow representations. E.g., [3] propose a deep unsupervised feature learning approach, but include only labeled data. The authors of [4] use stacked convo-

lutional autoencoders as well as independent component analysis with non-linearity for learning deep representations. He et al. [5] also use multi-layers of SR to obtain higher-level features. For this they combine a fully unsupervised feature learning procedure with hand-crafted feature extraction and pooling. The deep belief network of [6] benefits from the geometric data structure achieved by the local coordinate coding. They represent all data samples in two stacked layers as sparse linear combination of anchor points. However, so far all deep feature learning approaches do not produce fully interpretable representations.

In this paper, the overall goal is to learn deep features with the help of big amounts of unlabeled data, leading to good classification results and interpretable features, the latter being important for many classification or unmixing tasks [7]. We achieve this by designing a deep framework, called deep STL (DSTL), which combines STL and deep learning concepts. We extend the shallow approach of [8], which shows that high classification accuracies can be achieved by combining STL with archetypal dictionaries. Furthermore, we use the approach of [9] to find archetypes (extreme points of the data distribution), and adapt the dictionary learning to be suitable for our deep learning framework.

2. DEEP SELF-TAUGHT LEARNING

STL uses unlabeled data ${}^u\mathbf{X} = [{}^u\mathbf{x}_q], q = 1, \dots, Q$, training data ${}^{tr}\mathbf{X} = [{}^{tr}\mathbf{x}_n], n = 1, \dots, N$ with labels ${}^{tr}\mathbf{y} = [{}^{tr}y_n]$ given, and test data ${}^t\mathbf{X} = [{}^t\mathbf{x}_p], p = 1, \dots, P$, also with labels ${}^t\mathbf{y}$. All data samples consist of M -dimensional feature vectors $\mathbf{x} \in \mathbb{R}^M$, and the labels $y \in \{1, \dots, c, \dots, C\}$, where C is the number of classes. The labels are also represented by target vectors $\mathbf{t} = [t_c]$ of length C coding the label with $t_c = 1$ for $y = c$ and $t_c = 0$ otherwise.

We use SR approximating each sample by a linear combination of only a few elements of a dictionary. We initialize the dictionary following archetypal analysis [10], achieving a sparse data approximation $\mathbf{x}_n \approx \mathbf{D} \boldsymbol{\alpha}_n$, with an $(M \times K)$ -dimensional dictionary \mathbf{D} . To learn the dictionary $\mathbf{D} = [\mathbf{d}_k]$ we apply simplex volume maximization (SiVM) [9]. This approach finds archetypes as extreme points lying on the convex hull of the unlabeled data set $\{\mathbf{d}_k\} \in \{{}^u\mathbf{x}_q\}$, where $K \leq Q$. The coefficient vectors $\boldsymbol{\alpha}_n$ are the new SR of the data sam-

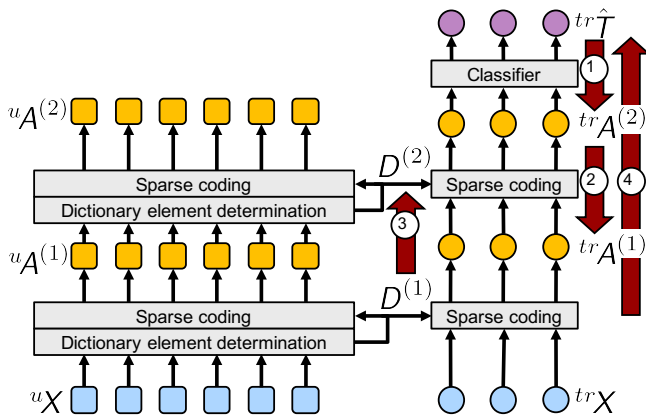


Fig. 1. Structure and update procedure of the DSTL approach: (left block) unlabeled and (right block) labeled data with classifier. The numbers represent the update procedure: ① gradient descent update of the L^{th} training representation; ② backpropagation of the gradient; ③ dictionary update with training representations; ④ learning new training representations and classifier.

ples. We derive these SRs by minimizing $\|D\alpha_n - x_n\|$ subject to non-negativity constraint $\alpha_{kn} \geq 0$ for all k and sum-to-one constraint $\sum_{k=1}^K \alpha_{kn} = 1$.

We extend the STL approach to DSTL to learn deep representations. Our network structure is shown in Fig. 1. The left side illustrates the exploitation of unlabeled data, and the right side the used of the labeled data; the data and their representations are symbolized by filled rectangle for the unlabeled and by circles for the labeled data. In general the network consists of L layer, where Fig. 1 shows the structure for $L = 2$. Pre-training is performed layer-wise, so that in each layer $l = 1, \dots, L$ the dictionary elements $D^{(l)}$ are determined from uX for $l = 1$ and from $uA^{(l-1)} = [u\alpha_q^{(l-1)}]$ for $l \neq 1$. Given the dictionaries, we learn $uA^{(l)}$ by least square estimation with non-negativity and sum-to-one constraint (left side of Fig. 1). Likewise, we learn $trA^{(l)}$ of the labeled data of the network from $D^{(l)}$. We train the logistic regression model ([11] p. 205-210) with the new features $trA^{(L)}$ predicting conditional probabilities $P(c|tr\alpha_n)$, which we can interpret as $tr\hat{T} = [tr\hat{t}_n]$.

To improve the network the reconstruction error can be minimized by updating the networks parameter. The number of parameters which need to be learned in a 2-layers network are $(M + K^{(2)}) \cdot K^{(1)} + (K^{(2)} + 1) \cdot C$ with $K^{(l)}$ being the number of dictionary elements in the layer (\cdot). It contains the number of dictionary entries in each layer and parameters of the classifier models. In order to learn these parameters we perform the following steps illustrated by ① - ④ in Fig. 1: In the first update step the dictionaries are fixed and only the training representations are updated. Given trt_n and $tr\hat{t}_n$, the backpropagation loss function is given by the following

equation:

$$J(tr\alpha_n^{(L)}) = \frac{1}{2} \|trt_n - tr\hat{t}_n\|^2. \quad (1)$$

Here the target vectors trt_n expresses the true membership of the training samples to the classes in the form of the 1-of-C coding scheme, and the $tr\hat{t}_n$ is the likewise encoded conditional probability for class membership estimated by our network. With the help of the gradient of this loss function with respect to the training representations $tr\alpha_n^{(L)}$ we update the training representations of the last layer:

$$tr\alpha_{kn}^{*(L)} = tr\alpha_{kn}^{(L)} + a \frac{\partial J(tr\alpha_n^{(L)})}{\partial tr\alpha_{kn}^{(L)}}. \quad (2)$$

Here the gradient is clipped element-wise to a threshold t_1 to avoid too excessive modifications of the representations [12]. We then backpropagate the gradient through the net in order to update all labeled representations using

$$trA^{*(l)} = D^{(l+1)} trA^{*(l+1)}, \quad (3)$$

for $l = L - 1, \dots, 1$ (① and ②).

In step ③, given the updated training representations, we update the dictionaries $D^{(l)}$ using the gradient descent method as proposed by [3]. To compute the dictionary update, we define a loss function

$$J_D(D^{(l)}) = \frac{1}{2} \|D^{(l)} trA^{*(l)} - trA^{*(l-1)}\|^2, \quad (4)$$

which will be minimized. In the first layer, $trA^{*(0)}$ is given by the original data trX . The gradient descent updating rule for the k^{th} dictionary element of the l^{th} layer is given by

$$d_k^{(l)} = d_k^{(l)} - \gamma (D^{(l)} trA^{*(l)} - trA^{*(l-1)}) tr\alpha_k^{(l)}, \quad (5)$$

where γ is the learning rate. Again the gradient is clipped to a threshold t_2 . Due to the dictionary updates their entries do not represent raw data samples anymore, which makes them not interpretable. We want to keep the dictionary elements interpretable by restricting them to true data samples. In case a dictionary element has changed sufficiently, we shift it to the nearest neighbor in feature space which contains the set of unlabeled data samples. Step ④ finally readjusts the labeled representations with the updated $D^{(l)}$ by minimizing the reconstruction error of trX and updates the classifier. We iterate steps ① - ④ until convergence of the dictionaries.

3. EXPERIMENTAL SETUP AND RESULTS

In this section we test our DSTL approach for two multi-spectral image data sets. For this we apply our two layered DSTL to the data sets and compare the accuracy with the results of a simple logistic regression. In Section 3.1 the two data sets are briefly introduced, followed by the data (Sec. 3.2) and experimental setup (Sec. 3.3). Finally, the results are presented in Sec. 3.4.

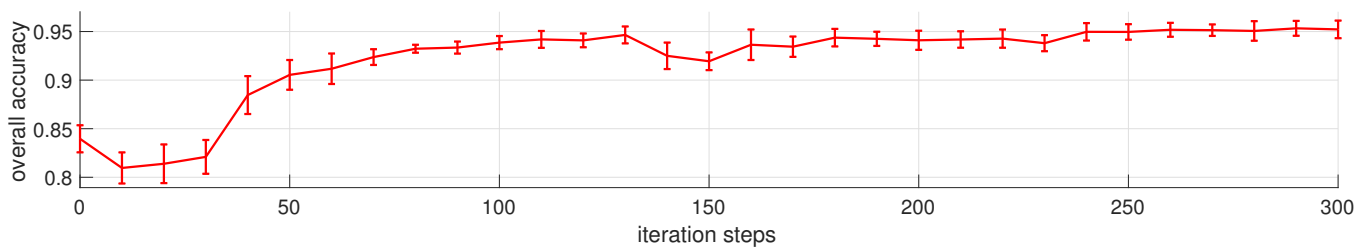


Fig. 2. Average and standard deviation over 10 runs of overall accuracy [%] of the DSTL approach for the Landsat 5 Data Set (right) over up to 300 iterations..

3.1. Data Sets

We use the following two multi-spectral data sets for the testing of our approach:

Landsat 5 Data Set: Our first data set is a multi-spectral Landsat 5 TM image from a study area located in the North of Novo Progresso (South America). It contains data for 6 bands (red, green, blue, Mid-Infrared and two NIR-bands) for $6,962 \times 7,921$ pixels. Parts of the image are labeled (approx. 57,000 pixels) with 6 classes (see Tab. 1). Additionally, we collect about 600,000 image patches from diverse areas worldwide as unlabeled data samples.

Zurich Summer Data Set: The Zurich Summer Data Set, provided by the University of Zurich [13], contains 20 VHR images of Zurich recorded 2002 by the QuickBird satellite. The images comprise four bands: red, green, blue and NIR with spatial resolution of 61.5 cm. They are labeled by 8 urban and periurban classes (roads, buildings, trees, grass, bare soil, water, railways, and swimming pools).

3.2. Data Setup

For both data sets we choose 5×5 -pixel image patches, leading to 100-dimensional input feature vectors for the Zurich Summer Data Set and 150-dimensional input feature vectors for the Landsat 5 Data Set. These input vectors are global contrast normalized and then shifted to positive values as input vectors for the DSTL network.

Landsat 5 Data Set: We randomly extract ten sub-data sets with 1,000 training samples (^{tr}X), around 56,000 test samples (^{t}X), and 1,000 validation samples each from the Landsat 5 image. We use the patches (around 600,000 pixels) as unlabeled samples (^{u}X).

Zurich Summer Data Set: We randomly extract 9,500 test samples (^{t}X) and 500 validation samples from one of the 20 images, and 1,000 training samples (^{tr}X) from the remaining 19 images. The 10,000 unlabeled (^{u}X) samples are randomly selected from all images. The data selection is done 20 times, so that the test and validation data are selected from each image.

3.3. Experimental Setup

In our experiments we create a DSTL with two layers to test if the test accuracy benefits from the DSTL approach over a simple logistic regression. We perform the following experiments on the two data sets:

Landsat 5 Data Set: The DSTL approach is carried through with 20 archetypes in the first layer and 30 in the second.

Zurich Summer Data Set: The DSTL approach is performed with 30 and 40 archetypes for the two layers.

In all experiments the threshold of the gradient clipping is set to $t_1 = t_2 = 0.001$ and the learning rates to $\alpha = \gamma = 1$. The DSTL update is iterated 1,000 times to find the best dictionary, judged by application to the validation data. We achieve the best results, in terms of classification accuracy, by stacking the representations of all layers for classification, similar to the idea used in denseNet [14].

3.4. Results

In all our experiments we achieve an improvement over the original representations:

Landsat 5 Data Set: Table 1 shows the class-wise, overall, and average test accuracy as well as the Kappa coefficient for the experiment of the Landsat 5 Data Set. Our DSTL approach with stacked representations yields better overall and average accuracies than the original data, but the Kappa coefficient is decreased. The average and standard deviation of the overall accuracy is illustrated in Fig. 2. It becomes obvious, that the average increases over the most iterations and the standard deviation is with maximal 0.03 % very small.

Zurich Summer Data Set: Table 2 shows that the DSTL with stacked features leads to an improvement in the mean over-all and mean average accuracy. Here also the kappa coefficient increases. Running this experiment with 1,000 iterations with Matlab version 16b on an Intel Core i5-2400 processor takes ca. 14 hours.

We expect to further improve the results by using larger dictionaries, but this will significantly increase run time.

Table 1. Class-wise accuracies [%], overall accuracy [%], average accuracy [%] and Kappa coefficient (Kappa) obtained by logistic regression using the original features of the Landsat 5 Data Set, and logistic regression on the stacked deep representations of the DSTL approach. The average results over ten runs is given with the standard deviation. The best results are highlighted in bold-print.

	original features	DSTL features
water	90.03 ± 23.06	97.21 ± 2.23
urban	88.90 ± 5.08	91.44 ± 4.94
secondary forest	55.43 ± 7.94	74.29 ± 3.33
pasture	99.11 ± 0.83	97.81 ± 1.54
burned pasture	100.00 ± 0.02	99.96 ± 0.08
primary forest	99.88 ± 0.07	99.34 ± 0.13
overall	94.23 ± 0.90	96.17 ± 0.37
average	88.89 ± 4.17	93.34 ± 1.00
Kappa	0.86 ± 0.02	0.79 ± 0.07

Table 2. Mean results of the logistic regression of the original Zurich Summer Data Set and of the stacked deep representations of the DSTL approach.

	original features	DSTL features
overall accuracy	68.0 %	74.1 %
average accuracy	60.1 %	65.9 %
Kappa	0.54	0.55

4. CONCLUSION

In this work we present a deep self-taught learning framework to combine the advantages of the STL with interpretable dictionaries and the deepness of neural networks. The accuracy is tested with two different multi-spectral image data sets. Further research will deal with a deeper DSTL network and interpretable non-linearity to raise the the accuracy. Altogether, the deep self-taught learning framework profits by the huge amount of unlabeled data, so that the learned deep features improve the results compared to classification results using the original feature representation and are achieved with still interpretable dictionaries.

Acknowledgments

This work has partly been supported by the EC under contract number H2020-ICT-644227-FLOURISH.

5. REFERENCES

- [1] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE*, 2013.
- [2] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, “Self-taught learning: transfer learning from unlabeled data,” in *ICML*, 2007.
- [3] Y. Gwon, M. Cha, and HT Kung, “Deep sparse-coded network (dsn),” in *ICPR*, 2016.
- [4] R. Kemker and C. Kanan, “Self-taught feature learning for hyperspectral image classification,” *IEEE*, 2017.
- [5] K. He, Y. and Kavukcuoglu, Y. Wang, A. Szlam, and Y. Qi, “Unsupervised feature learning by deep sparse coding,” in *SIAM*, 2014.
- [6] Y. Lin, T. Zhang, S Zhu, and K. Yu, “Deep coding network,” in *NIPS*, 2010.
- [7] C. Römer et al., “Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis,” *Functional Plant Biology*, 2012.
- [8] R. Roscher, C. Römer, B. Waske, and L. Plümer, “Land-cover classification with self-taught learning on archetypal dictionaries,” in *IGARSS*, 2015.
- [9] C. Thureau, K. Kersting, and C. Bauckhage, “Yes we can: simplex volume maximization for descriptive web-scale matrix factorization,” in *CIKM*, 2010.
- [10] A. Cutler and L. Breiman, “Archetypal analysis,” *Technometrics*, 1994.
- [11] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016, <http://www.deeplearningbook.org>.
- [13] M. Volpi and V. Ferrari, “Semantic segmentation of urban scenes by learning local class interactions,” in *CVPR*, 2015.
- [14] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely connected convolutional networks,” *arXiv preprint arXiv:1608.06993*, 2016.

FORECASTING IONOSPHERIC TOTAL ELECTRON CONTENT MAPS WITH DEEP NEURAL NETWORKS

Noëlie Cherrier, Thibaut Castaings, Alexandre Boulch

ONERA, The French Aerospace Lab,
Chemin de la Hunière, 91123 Palaiseau, France

ABSTRACT

Satellite telecommunications and Global Navigation Satellite Systems (GNSS) would benefit from an early prediction of the ionospheric activity. The Total Electron Content (TEC) values of the ionosphere are already locally predicted by models from previous studies, but no model exists to our knowledge for worldwide prediction. A large amount of data for world TEC maps is available from the Center for Orbit Determination in Europe (CODE). With Deep Neural Networks (DNN), we propose a method to forecast a sequence of global TEC maps following past given TEC maps, without introducing any prior knowledge. By combining several state-of-the-art architectures, the proposed approach is competitive with previous works on TEC forecast, while predicting global TEC maps.

Index Terms— Ionosphere, TEC, forecast, deep learning, neural networks, sequence prediction

1. INTRODUCTION

Satellite telecommunication services and Global Navigation Satellite Systems (GNSS) are widely used services subject to perturbation due to the ionospheric activity. During high ionospheric activity, the path of transionospheric radio waves indeed changes, inducing significant bitrate reduction and positioning errors [1, 2]. As a consequence, forecasting the ionosphere state globally (*i.e.* worldwide) increases the ability of the users to evaluate, for example, data loss probabilities or margin of error in positioning planning.

The ionospheric activity is usually measured using Total Electron Content (TEC), which is the total number of electrons in the ionosphere integrated along a vertical path above a given location. It is expressed in TEC Units ($1 \text{ TECU} = 10^{16} \text{ el}/\text{m}^2$), usually ranging from a few units to one hundred TECU.

Several services exist to address TEC forecasting. They rely on measurements provided by GNSS ground networks [3] and aim at producing global TEC maps. CTIPE is an experimental tool implementing complex physics models [4] developed by the US Space Weather Prediction Center that produces global forecasts 30 minutes ahead of real-time. In

Europe, the ESA Ionospheric Weather Expert Service Center combines products from different national services to provide global and regional 1-hour TEC forecasts. However, the records of the input data and forecasts are not published.

A global analytical TEC model has been proposed in [5], using open source TEC data from the Center for Orbit Determination in Europe (CODE). This model is intended to apply to any temporal range, without relying on a record of TEC values.

The literature provides several methods using time series and statistical methods to predict TEC with various forecasting horizons from a few minutes to several days based on the previous state of the ionosphere. Most of these methods [6, 7, 8, 9, 10] provide predictions above specific stations. Among these, a few works aim at reconstructing the TEC on a small area [11, 12] with methods such as Bezier surface-fitting or Kriging. Some of them use machine learning, particularly neural networks [11, 13, 14], to infer the model parameters. However, they only focus on local stations and obtaining a regional or global prediction would require one model for each location and interpolation for not covered areas.

In this paper, we aim at predicting global TEC maps from 2 to 48 hours ahead of real-time. We propose a purely data-driven approach using deep convolutional recurrent networks. Deep Neural Networks (DNN) have the advantage to enable complex modeling of large input data, such as global TEC maps in this case, with little or no prior knowledge.

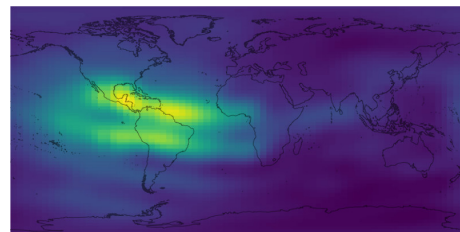
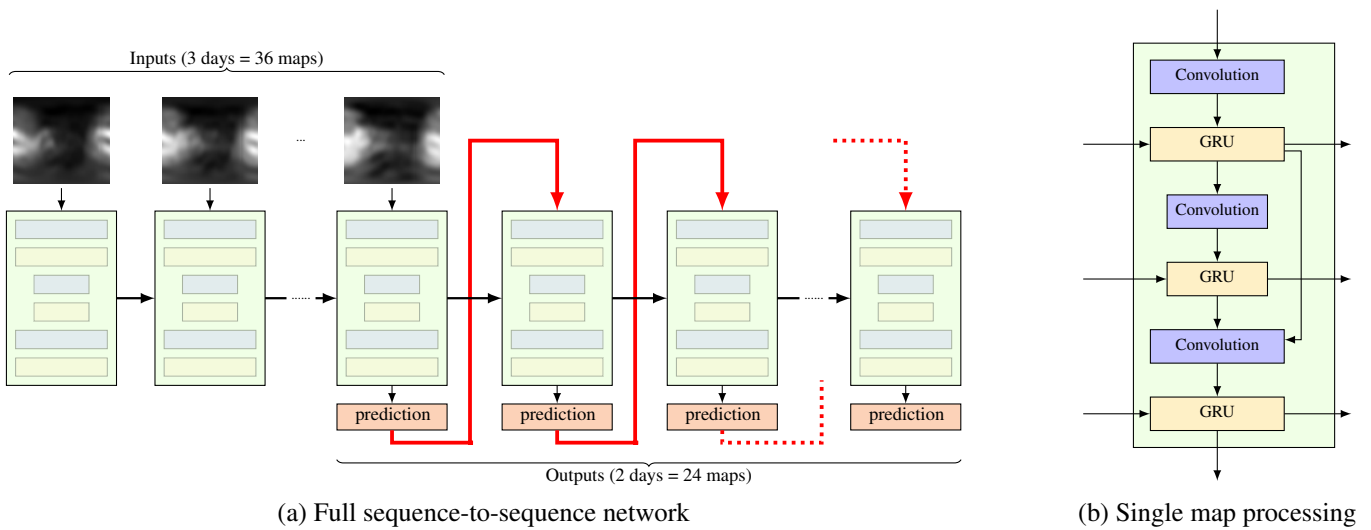


Fig. 1. TEC map example

This work is a supplementary study following the works published in [15] and investigates the possible improvement provided by an alternative neural network architecture. The paper is organized as follows: Section 2 presents the recurrent


Fig. 2. Network architecture

U-net architecture and Section 3 shows quantitative results on TEC prediction.

2. METHOD

A neural network architecture is designed considering that we want to output a sequence of 48 hours of TEC maps given a number of past maps.

2.1. Data retrieving and preprocessing

Open source TEC data from the CODE is used in this study, the TEC maps having a $5^\circ \times 2.5^\circ$ resolution on longitude and latitude and 2-hour temporal resolution, covering all latitudes and longitudes (Fig. 1). One pixel in these maps represents the vertical TEC at this point. The study is conducted using the data from 1/1/2014 to 12/31/2016.

A 24-hour periodicity can be easily noticed while observing the data, due to the Sun heating the ionosphere during the day. There is no interest of having a neural network learn deterministic phenomena such as the day-night cycle. As in [15], we make the effect of Earth's rotation no longer visible to the neural network by changing the frame of reference to Heliocentric. Finally, the data is loaded as a sequence of 60 maps (one map every two hours): the first 36 maps (*i.e.* 3 days) are fed to the network, the last 24 maps (*i.e.* 48 hours) being the prediction targets.

2.2. Network architecture

The challenge of this study is to design a neural network able to handle a specific sequence prediction problem in which both the inputs and targets are a sequence of images (*i.e.* TEC maps).

Global architecture. The underlying idea of this paper is that a large part of future ionospheric activity can be inferred from its previous states. Particularly when looking at the temporal evolution of TEC maps, the main phenomena are continuous, which supports the possibility of predicting the next map sequence. This temporal trend is extracted via Recurrent Neural Networks (RNN), allowing temporal information to flow between processed maps and assist the prediction. The whole pipeline is presented in Fig. 2 (a). The sequence is processed frame by frame in the temporal order by a recurrent convolutional neural network (green block in Fig. 2 (a)), the temporal information being kept during the iterations of the process. The prediction process is achieved by recursively feeding the next column of the network with the last prediction (red arrows).

Computational block. The recurrent convolutional neural network used to process one temporal frame is presented in Fig. 2 (b). Convolutional Neural Networks (CNN) are used to handle the bidimensional structure of the TEC maps. Also, as we need to output TEC maps, an architecture similar to U-net [16] is exploited as an alternation of convolutional layers and recurrent units. Three Gated Recurrent Units (GRU) [17], a special kind of Recurrent Neural Network (RNN) are used to capture the temporal dependencies at different spatial scales.

Training For each predicted map, the individual cost function is defined as the sum of the relative error and the ℓ_1 loss with respect to the ground truth. The final cost function is summed over the maps produced by the successive prediction columns.

3. RESULTS

Once it has been trained, the network can be fed with any 3-day sequence from the test set and produce 2-day forecasts consecutive to this sequence. The Root Mean Square (RMS) error (1) is used to assess the performance of the model.

$$\epsilon_{RMS} = \sqrt{\sum_{t \in \mathcal{S}} \sum_{i \in \mathcal{M}^t} (P_i^t - T_i^t)^2} \quad (1)$$

with \mathcal{S} the sequence of TEC maps, \mathcal{M}^t the TEC map at t , P the predicted map and T the ground-truth map, where t indexes time and i is the map pixel index.

3.1. Comparison on sequence prediction

For benchmarking purposes, we set up three reference methods. The first one is a basic constant prediction: the mean over the input sequence. The next one is called periodic prediction, the predicted sequence is exactly the input of the last two days. Finally, the approach proposed in [15] is a less complex neural network relying on Long Short-Term Memory (LSTM) networks (an other type of RNN) and CNN. The overall data flow (Fig. 2 (a)) is similar, but the single map processing block (Fig. 2 (b)) is implemented by an Encoder-LSTM-Decoder cell.

Table 1 presents the results of our experiments. We evaluate the performance of our architecture against the three baseline models. We also add in the table the scores of our model where we replaced the convolutional GRU units by LSTM ones.

First column presents the best scores obtained over several trainings. The approach from [15] gets a mean RMS of 2.407 TEC units (the average TEC value being around 30 TEC units) while the proposed network RMS averaged over the predicted sequence is 2.373 TEC units. Comparing to periodic prediction and convolutional LSTM [15], the performance is improved by respectively more than 8% and more than 1% using the U-net and GRU cell.

To our interpretation, this improvement comes from the higher interdependence between recurrent maps. Our network uses three recurrent units against one in [15]. The temporal behavior is captured at different spatial scales. Particularly, details do not suffer from the high compression rate operated by the encoder in [15].

However, as shown in the second column, these networks have difficulties to converge and on several trainings they do not reach a satisfactory performance. The bad scores are mainly obtained in the 24-48 hours prediction range. As underlined by the third column (first 24 hours mean RMS errors), our networks perform very well for the first 24 hours predictions, overcoming [15] by 0.1 TEC unit. We understand the difficulty to predict the 24-48 hours range as a long-term dependency only understood by a few trainings (the best runs,

Table 1. Comparison with other methods. RMS are expressed in TEC units

Method	RMS 48h (best)	RMS 48h (mean)	RMS 24h (mean)
Constant	3.18	3.18	3.12
Periodic	2.59	2.59	2.59
LSTM [15]	2.407	2.69	2.56
Ours LSTM	2.38	2.67	2.45
Ours GRU	2.373	2.69	2.43

that outperform periodic prediction on the whole 2-48 hours range).

3.2. Comparison with literature

In Table 2, we compare the results for the proposed approach with results from state-of-the-art models. The presented RMS errors are computed by selecting the same latitude(s) of the station(s) studied in the cited paper, as well as the same period of study.

Table 2. Results of previous works

Reference	RMS (ref)	RMS (proposed)
[6] Chunli D., Jinsong P.	1.45	2.049
[13] Huang, Z., Yuan, H.	≤ 2	1.936
[9] Niu, R. <i>et al.</i>	3.1	0.800

The obtained results are competitive with state-of-the-art models (their RMS errors range from 1.5 to 3 TEC units) and the proposed approach provides global TEC map forecasting 2 to 48 hours ahead of real-time. However, the comparison with previous works on TEC forecast is only indicative since these works differ by their prediction horizons and since several studies focus on one or a few specific measuring stations instead of producing a worldwide TEC prediction.

4. CONCLUSION AND PERSPECTIVES

In this work, we extend the method of [15] for TEC sequence prediction given the previous TEC maps. By investigating a new network architecture based on multiple recurrent neural units, we show that using more interlinked spatial processing (convolutional layers) and temporal information passing (recurrent units) leads to improved results.

Among the future work, the convergence issues requiring several training will be investigated. We will also explore the inclusion of more input information. There are complex dependencies that may not depend on the previous states of the ionosphere. As an example solar particles may interfere with the ionosphere [18, 19]. The next step will consist in using

solar activity as an additional input to the network. Several information sources are considered such as multispectral solar images or solar wind.

5. REFERENCES

- [1] S. Datta-Barua, J. Lee, S. Pullen, M. Luo, A. Ene, D. Qiu, G. Zhang, and P. Enge, "Ionospheric threat parameterization for local area GPS-based aircraft landing systems," *Journal of Aircraft*, vol. 47, no. 4, pp. 1141–1151, 2010.
- [2] J. Lee, S. Datta-Barua, G. Zhang, S. Pullen, and P. Enge, "Observations of low-elevation ionospheric anomalies for ground-based augmentation of GNSS," *Radio Science*, vol. 46, no. 6, pp. 1–11, 2011.
- [3] E. Tulunay, E. T. Senalp, L. R. Cander, Y. K. Tulunay, A. H. Bilge, S. S. Kouris E. Mizrahi, and N. Jakowski, "Development of Algorithms and Software for Forecasting, Nowcasting and Variability of TEC," *Annals of Geophysics, Supplement to Volume 47*, vol. 47, no. 2/3, pp. 1201–1214, 2004.
- [4] G. H. Millward, I. C. F. Muller-Wodarg, A. D. Aylward, T. J. Fuller-Rowell, A. D. Richmond, and R. J. Moffett, "An investigation into the influence of tidal forcing on F region equatorial vertical ion drift using a global ionosphere-thermosphere model with coupled electrodynamics," *Journal of Geophysical Research: Space Physics*, vol. 106, no. A11, pp. 24733–24744, 2001.
- [5] N. Jakowski, M. M. Hoque, and C. Mayer, "A new global TEC model for estimating transionospheric radio wave propagation errors," *Journal of Geodesy*, vol. 85, no. 12, pp. 965–974, 2011.
- [6] D. Chunli and P. Jinsong, "Modeling and prediction of TEC in China region for satellite navigation," in *2009 15th Asia-Pacific Conference on Communications*, Oct 2009, pp. 310–313.
- [7] N. A. Elmumim, M. Abdullah, and A. M. Hasbi, "Improving ionospheric forecasting using statistical method for accurate GPS positioning over Malaysia," in *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEEES)*, Nov 2016, pp. 352–355.
- [8] X. Li and D. Guo, "Modeling and prediction of ionospheric total electron content by time series analysis," in *2010 2nd International Conference on Advanced Computer Control*, March 2010, vol. 2, pp. 375–379.
- [9] R. Niu, C. Guo, Y. Zhang, L. He, and Y. Mao, "Study of ionospheric TEC short-term forecast model based on combination method," in *2014 12th International Conference on Signal Processing (ICSP)*, Oct 2014, pp. 2426–2430.
- [10] X. Zhenzhong, W. Weimin, and W. Bo, "Ionosphere TEC prediction based on Chaos," in *ISAPE2012*, Oct 2012, pp. 458–460.
- [11] Ersin Tulunay, Erdem Turker Senalp, Sandro Maria Radicella, and Yurdanur Tulunay, "Forecasting total electron content maps by neural network technique," *Radio Science*, vol. 41, no. 4, 2006.
- [12] Y. W. Wu, R. Y. Liu, Wang Jian-Ping, and Z. S. Wu, "Ionospheric tec short-term forecasting in china," in *Proceedings of the 9th International Symposium on Antennas, Propagation and EM Theory*, Nov 2010, pp. 418–421.
- [13] Z. Huang and H. Yuan, "Ionospheric single-station TEC short-term forecast using RBF neural network," *Radio Science*, vol. 49, no. 4, pp. 283–292, 2014.
- [14] Erdem Turker Senalp, Ersin Tulunay, and Yurdanur Tulunay, "Total electron content (TEC) forecasting by Cascade Modeling: A possible alternative to the IRI-2001," *Radio Science*, vol. 43, no. 4, 2008.
- [15] N. Cherrier, T. Castaings, and A. Boulch, "Deep sequence-to-sequence neural networks for ionospheric activity map prediction," in *24th International Conference On Neural Information Processing (ICONIP)*, November 2017.
- [16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *CoRR*, vol. abs/1505.04597, 2015.
- [17] K.Cho, B. Van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *CoRR*, vol. abs/1406.1078, 2014.
- [18] D. F. Webb, "Coronal mass ejections: origins, evolution, and role in space weather," *IEEE Transactions on Plasma Science*, vol. 28, no. 6, pp. 1795–1806, Dec 2000.
- [19] H. W. Wells, "Effects of Solar Activity on the Ionosphere and Radio Communications," *Proceedings of the IRE*, vol. 31, no. 4, pp. 147–157, April 1943.

CROSS-MATCH OF ASTROMETRIC CATALOGUES PERFORMED WITH DATABASE SPATIAL TECHNOLOGIES

A. F. Mulone¹, R. Morbidelli², R. Messineo¹, M. Lattanzi², R. De Marchi¹, A. Vecchiato²

¹ALTEC S.p.A., Corso Marche 79, 10146 Torino, Italy <http://www.altecspace.it>

²INAF-OATo, Via Osservatorio 20, 10025 Pino Torinese, Torino, Italy <http://www.oato.inaf.it>

ABSTRACT

Data Processing Centre of Turin (DPCT) collects and maintains data of Gaia Mission. An important subset of these data is represented by the astrometric catalogues released during the Gaia mission. All collected data, including some past astronomical catalogues, are kept online in order to be available to DPCT scientists for sky-oriented analyses.

The availability of data has stimulated the cross-match of Gaia catalogues with different external catalogues. DPCT uses Oracle as database technology and Oracle Spatial features were enabled to permit spatial-criteria-based catalogues cross-match, which did not alter nominal structures of data stored in DPCT databases [1]. New dedicated data structures were defined instead.

Enabling spatial cross-match was possible thanks to the insertion of a new reference system in Oracle database: the International Celestial Reference System (ICRS) that is widely used for astronomical catalogues but it is not present in commercial databases [2][3]. Moreover developing ad-hoc store procedure to execute big catalogues crossmatch was necessary to reach results.

Index Terms— Cross-Match, Catalogue, Big Data, Spatial, Oracle, Data Management, Indexing, DPCT, GAIA

1. INTRODUCTION

The cross-match between two or more astronomical catalogues needs a unique reference system for the sky coordinates of objects. Many reference systems are possible, but ICRS system (Fig. 1) is the most used, it is indeed also adopted in Gaia.

Thanks to the collaboration between scientists and technicians, ICRS system was defined and implemented in Oracle databases. The measurement unit used in the model is the radian instead of the degree, as done in Gaia. However the support of ICRS system is not perfect, some Oracle service requests have been opened and they are still open because Spatial-development team intervention is required.

THE INTERNATIONAL CELESTIAL REFERENCE SYSTEM (ICRS)

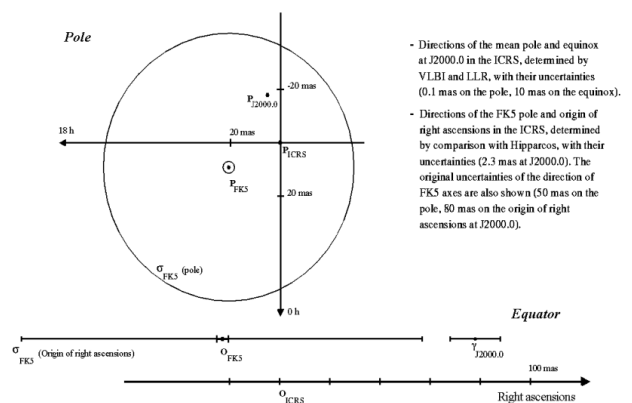


Figure 1: Directions and mean poles in ICRS and FK5 [2][3].

Before executing cross-matching, the resources usage is checked because extracted data are stored in a customized database schema accessible by scientists to execute their analyses. The resource consumption must be monitored because Oracle spatial research is based on objects, procedures and functions, so it has a cost to pay attention to.

2. THE ASTROMETRIC CROSS-MATCH

Astronomical catalogues, i.e. lists of astronomical objects, of the latest generation frequently cover most, if not all, of the celestial sphere and contain from one billion to several billion sources, as the detection limit of the observational surveys (either ground-based or space-borne) they were derived from is set to increasingly lower fluxes (fainter magnitudes). These numbers are huge compared to the best of similar catalogues from only 10-15 years ago [3]. Moreover, these compilations originate from processing elementary observations taken throughout the years of the operational lifetimes of such surveys, making the total number of entries increase by two orders of magnitude or more that of the sources itself.

The astronomical compilations of our interest usually provide, for each of the listed sources, coordinates and their time derivatives (i.e., proper motions) so that one can determine the direction on the celestial sphere to any of the entries as a function of time. The Gaia catalogue is the most important of these catalogues derived from space

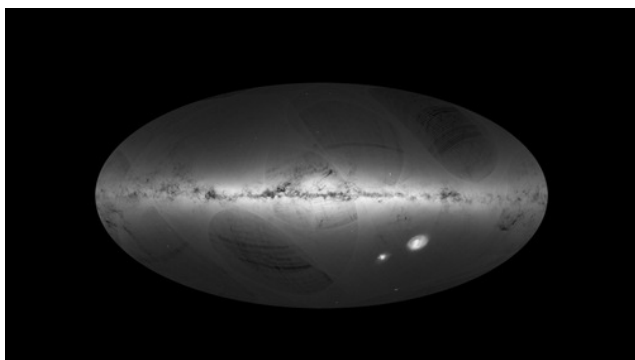


Figure 2: The Gaia DR1 catalogue of more than 1 billion stars.

observations [5], while examples of ground based analogues can be found in [4].

We also recall that the reference frame in which the astronomical coordinates (and their derivatives) are expressed does not correspond to any of the ones present in commercial databases. In fact, astronomical catalogues are usually in the ICRS Reference Frame [4] that, for example, has its Cartesian origin at the barycentre of the Solar System instead of, e.g., at the geocenter, as in the case of Earth referenced observations.

We can now appreciate what astrometric cross-matching is and why it can be critically important in astrophysics.

Astrometric cross-matching uses exclusively coordinates to find common sources (within some specified angular thresholds) in the cross-matched catalogues, despite the fact that flux and colour information might be listed as well. This is of particular importance, e.g., when matching multi-wavelengths astronomical catalogues (i.e., compiled from observations taken in different bands of the electromagnetic spectrum) as photons of different wavelengths might come from different regions within a source star. Therefore, by not making any assumption on the regions emitting the different colours, astrometric cross-matching might be instrumental in characterizing the physical nature of complex sources as the case of Seyfert galaxies [6].

3. SPATIAL TECHNOLOGY

All DPCT catalogues are stored in Oracle 12c databases that include a wide range of spatial analysis functions and services to evaluate how near or far an object is to another (Oracle Spatial and Graph requires the Enterprise Edition of Oracle Database). This is useful to check if some entity is present inside a particular region, or to visualize geospatial patterns on maps and imagery (Fig.3).

Oracle architecture includes partitioning, which consists in splitting a single logical table and its indexes into one or more physical tables, each one with its own index.

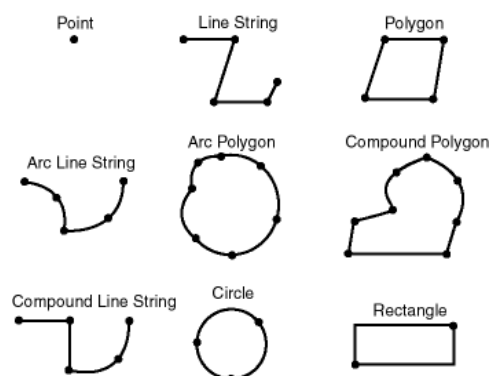


Figure 3: Geometric Types.

Spatial indexes associated with partitioned tables can be partitioned as well; Oracle allows creating partitioned spatial indexes only if the table associated with the index is partitioned as well. Moreover, the only partitioning scheme supported for spatial indexes is the partitioning by range.

The creation of spatial R-tree indexes and graph B-tree indexes can be split into smaller tasks that can be performed in parallel. Using free hardware (CPU) resources can substantially increase performance and provides a significant time saving, obviously depending on the spatial datasets, the index types and parameters. Spatial queries can run in parallel on partitioned spatial indexes, improving the performance of "within distance", "nearest neighbour", and "relate" queries. Performance scales with the number of CPUs used to execute a query.

Spatial and Graph uses a two-tier query model to resolve spatial queries:

- the primary filter
- secondary filter

Primary filter permits fast selection of candidate records to pass along to the secondary filter.

Secondary Filter applies exact computations to geometries that result from the primary filter. In some cases primary filter only is executed. Spatial index implement the primary filter.

4. DATA ORGANIZATION

4.1. Reference system

ICRS model is used by all catalogues of Gaia mission and for every external catalogue scientists provide to DPCT. Without the introduction in DPCT databases of the ICRS reference system there would be no possibility to enable spatial indexing because Oracle has not implemented any reference systems for space data.

ICRS was modelled exploiting the ellipsoid model, already defined in Oracle 12c. Two axes are right ascension and north declination while the inverse flattening parameter is set to null. After introducing astronomical reference system it was possible to create spatial data structures.

A new table was created for each spatial-enabled catalogue to preserve Gaia data model. Spatial tables contain the following fields:

- element identifier;
- radian coordinates (ascension and declination);
- spatial object field (SDO_GEOMETRY).

The spatial object is built specifying its ICRS reference system, its geometry (each catalogue element is modelled as a 2-D point) and its coordinates.

4.2. Spatial enabling

The spatial index is required to enable the spatial research on the SDO_GEOMETRY field. In order to build the index, information about geometry field name, reference system used and axes must be inserted in a system table, defined for each database user, where the spatial is enabled. In spatial index creation clause you must specify 'MDSYS.SPATIAL_INDEX'.

All these steps are required for each ingested catalogue where we want to perform spatial researches. Through a spatial index, you can search for all the elements of the catalogue within a specified distance from a fixed point. This is the most common type of query executed during catalogue cross-match.

Spatial structures have a remarkable weight in disk space management: the SDO_GEOMETRY field of the spatial table (which has only three numeric fields' identifier, ascension and declination, in addition to the spatial one) weighs almost half of the table total. Another important aspect to be considered is that the index space consumption is about 80% of the weight of the 'spatial table' and it is almost three times bigger than the weight of the partitioned index on the catalogue element identifier.

Time for indexes creation is very high when you have billions of row and many times the index creation process could fail because there is saturation of system tablespaces and the index creation could exceed the undo retention of database. To avoid problem of this type you must create a partitioned index in unusable way.

Creating a spatial partitioned index in unusable way you create structures of the indexes and you can proceed to build each index partitions. Moreover you can build in parallel different index partitions. In this way you avoid undo retention problems or system tablespaces saturation and you can create index in less time because you can proceed in parallel. Spatial indexes can only be partitioned by range.

To have a spatial range partitioned index you must create it on a table partitioned by range. It is not important the field used for partitioning, it cannot be related to spatial coordinates fields. Creating a table partitioned by range you can declare the spatial index as local. This is what was done at DPCT.

Selecting a point within a specified radius is achievable thanks to the function 'SDO_WITHIN_DISTANCE'. You

must pass to this function two points, the distance value and it return TRUE if distance is less or equal to the passed value or FALSE in other case. This is the main function used to execute crossmatch at DPCT. If one of the point that must be used it is external to tables involved in the query you have to create a new SDO_GEOMETRY object (this consideration is valid for any geometry object not for point only).

Measuring distance between two geometric objects (which are 2-D points at DPCT) is executed by function 'SDO_GEOM.SDO_DISTANCE' that takes two geometric objects and a tolerance value. Tolerance is used to associate a level of precision with spatial data, it reflects the distance that two points can be apart and still be considered the same. The tolerance value must be a positive number greater than zero. An optional parameter that could be used is the unit of measurement but as you can see in the next paragraph this is not applicable using ICRS system defined in Oracle until now.

4.3. Issues found

The searching and distance computation works on ICRS reference system, but we found some bugs related to the distance calculus in particular.

During the first attempt to define ICRS model we used a 3-D ellipsoid. This model worked and it permitted to find all the elements within a certain distance from a star but it did not allow computing distances between the star and the found objects. ICRS was defined as a 2-D model to solve this problem. This issue on ICRS model with a 3-D ellipsoid was found on Spatial implemented in Oracle 12.1 instead, Spatial in Oracle 11.2 was not affected by this problem.

Another bug we found occurs in the computation of the distance between two catalogue elements. The distance value is numerically correct, but Oracle database considers it expressed in metric measurement unit instead of angular, even if in ICRS the measurement unit of each model element is radian. While this is not a relevant bug as much as regards scientific analyses (numeric values are correct), the problem arising here is the conversion of the returned value to different angular units using the same function that computes the distances.

Finally, the cross-match of big catalogues (billions rows both in the input catalogue and in the catalogue where the spatial research is executed) has low performances exploiting only the database tools, so you should consider the possibility to create a specific software that uses database spatial features. The spatial research is based on spatial objects not on field consisting of basic type so you can spent a lot of time for your queries. You should decompose your big queries in many little queries that can be executed in parallel by a software.

5. USE CASES

A cross-match exercise executed at DPCT was the ‘Astrometric verification of kinematics substructures’. An input list of 2417 elements was provided and all the stars within a distance of 2” from each input were supposed to be provided to scientists.

To execute the cross-matching, we ingested into the database the input list with all the coordinates converted to radians and then we exploited spatial structures (on both table and index) built on about-3.5B-rows Gaia DR02 mission catalogue.

Thanks to indexing, we were able to identify every element within a radius of 2” (i.e., 0.00001 radians) from each input point ordering the found objects by distance from the input star.

The complete cross-match took about 4 hours, while searching for each element of the input catalogue in the Gaia catalogue takes 1 to 3 seconds (the only research took about 2.5 hours). After that, we had to execute a separate function to compute distances between every found element and the input star, because Oracle does not allow using a unique function for both research and distance computation, as said in the previous section.

Reported cross-match can be considered very light but more challenging was the cross-match between a catalog with about 1,3 millions of rows and a catalog with 3.5 billions of rows (the same catalog used in previous crossmatch).

To achieve a result in reasonable time was not possible to develop a stored procedure to execute crossmatch considering the whole catalogues not in a one only operation but to execute this crossmatch was necessary considering one of the catalogues divided in different partitions in order to execute different cross-match executing store-procedure in parallel on different partitions of one catalogues. The second catalog was considered in its entirety.

This case shows spatial research is not a soft activity, as well as the fact that searching is performed on objects heavier than the common datatypes like number, char, and so on.

Performances are very different considering the same catalog with 3.5 billions of rows cross-matched one time with a catalog of about 2 thousands of rows and another time with a catalog with 1.3 millions of rows. If you should execute cross-matching between two catalogues both with many billions of rows you could not face it using the same database store-procedures mentioned above but this type of experiment will requires a more complex and distributed application to reach the expected results and you should be able to adopt a different approach in some circumstances.

In the following table are reported time spent during cross-match among different number of rows of a little catalog with a catalog with 3.5 billions of rows. Test was executed for different radius size. Between brackets are reported elements found.

		<i>Radius</i>		
		$2*10^{-4}$ rad	$2*10^{-5}$ rad	$2*10^{-6}$ rad
Rows	1	1.63s (27 rows)	2.67s (7 rows)	1.40s (1 Row)
	10	19.06s (861 rows)	12.81s (29 rows)	12.8s (9 Rows)
	100	161.41s (7840 rows)	129.5s (387 rows)	127.27s (99 Rows)
	1000	1392.72s (49615 rows)	1265.11s (5499 rows)	1275.7s (1900 Rows)

6. FUTURE PROSPECTS

In the future other implementations of spatial technology provided in other RDBMSs will be tested at DPCT, in particular, solutions that support Astronomical reference system in order to face space consumption and spatial query performances.

At ESAC (Madrid), among different technologies, Gaia Archive is based on PstgreSQL instances with modules pgSphere and Q3C that allow capabilities for geometrical queries and crossmatch of astronomical catalogues.

A comparison between these different technologies should be a useful exercise.

7. ACKNOWLEDGMENTS

The authors are members of the Gaia Data Processing and Analysis Consortium (DPAC) and this work has been supported by the ASI (Italian Space Agency) contracts n. 2016-17-I.0 and n. I/058/10/0.

8. REFERENCES

- [1] Oracle 12c Spatial and Graph developer’s guide <https://docs.oracle.com/database/121/SPATL/toc.htm>
- [2] The International Celestial Reference System (ICRS) <https://www.iers.org/iers/en/science/ICRS/ICRS.html>
- [3] International Celestial Reference System (ICRS) http://aa.usno.navy.mil/faq/docs/ICRS_doc.php
- [4] Qi, Z., Yu, Y., Bucciarelli, B., et al., “Absolute Proper Motions Outside the Plane (APOP)—A Step toward the GSC2.4”, *AJ*, 150, 137, 2015.
- [5] Prusti, T., de Bruijne, J. H. J., et al., “Gaia Collaboration”, *AA*, 595, A1, 2016.
- [6] Lattanzi, M. G., “Astrometric Cosmology”, *Mem.S.A.It.*, 83, 1033, 2012.
- [7] Mulone A., Morbidelli R., Messineo R., De March R., Filippi F., Vaschetto M., Sella F., Uzzi S., Manetta C., “Data Acces Services at DPCT to support Scientists’ online and offline data analysis” in *Proc. Big Data from Space (BiDS16)*, IEEE, 2016

SERVING CONTINUOUS AND GLOBAL HIGH RESOLUTION SATELLITE DATA – AN EXAMPLE BASED ON SENTINEL-2 DATA

Rouven Volkmann, Christian Strobl, André Twele, Torsten Heinen, Christoph Reck

German Aerospace Center (DLR)
Earth Observation Center (EOC)
Oberpfaffenhofen, Germany

ABSTRACT

For the Copernicus Data and Exploitation Platform of Germany (CODE-DE) [1] a full resolution imagery service of Sentinel-2 Level-1C data has been developed. Based on the experiences of this project this paper provides a general overview on how OGC-compliant online access services of large earth observation data sets can be implemented.

Different means of implementation are compared through criteria such as the performance of the web services, the storage requirements and the processing power needed as well as the quality and flexibility reached. Considerations like projection, transparency, bit depth, formats and compression play a big role in serving the data in a performant way.

This work aims to suggest best practices and to assist on which option to choose for which dedicated use case.

Index Terms— Earth Observation data, Web Mapping Service, Web Coverage Service

1. INTRODUCTION

More than 4000 Sentinel-2A Level 1C (L1C) products are produced daily, each with a size of approximately 500 Megabytes. This results in more than 50 Terabytes per month or 600 Terabytes each year. Serving this amount of globally and continuously acquired high resolution satellite data in full resolution is a challenging task. Keeping several years of data online with instant availability requires large and fast storage systems. When storage capacities are limited, a rolling archive with automated eviction mechanisms is needed.

The original Sentinel-2 L1C data are available in zipped SAFE file format. Each zip file contains the imagery from one of the 100 x 100 kilometers grids of the Military Grid Reference System (MGRS) [2]. It contains the image data, meta-data, quality indicators (e.g. defective pixels masks) and auxiliary data [3]. The image data are provided with JPEG2000 lossless compression resulting in a file size of approximately 50% compared to conventional lossless compression methods. While this is beneficial for storing and providing the raw data, it has negative implications on rendering the imagery. Results from our performance tests show that reading

and rendering JPEG2000 compressed data is time consuming and requires much processing power. Although this is acceptable for nonrecurring tasks where small file sizes and good quality outweigh performance needs, it may be impracticable for continuous on-the-fly operations. Additionally, the free and open JPEG2000 image library OpenJPEG does not meet the performance requirements and proprietary drivers may be necessary.

As a consequence to these limitations, alternative implementations have been evaluated.

2. IMPLEMENTATION

2.1. Software

For the CODE-DE project, DLR chose to use free and open source software to enable the possibility for improvements during project runtime. This can be achieved through cooperations with the open source community.

The Geospatial Data Abstraction Library (GDAL) [4] is used for the geographic data processing needs and for the conversion of the original Sentinel-2 raster data stored in JPEG2000 format to the widely supported GeoTiff container format. GDAL can reproject the image data as well as adding supplementary information to the GeoTiff like transparent pixels, internal tiling and overviews. Overviews are one or several previews of the original image with reduced resolution to provide fast access when viewing in small scales. Internal tiling is necessary for an improved performance on larger scales, when only a small portion of the image needs to be loaded. The original Sentinel-2 L1C data include both overviews and internal tiling.

For rendering and serving the data, GeoServer is used. GeoServer [5] is a web-server based on Apache Tomcat for rendering and serving geographic data using the Open Geospatial Consortium (OGC) Web Mapping (WMS) and Web Coverage service (WCS) standards [6]. In this use context, the most important new feature in GeoServer is the possibility to add granules of different coordinate reference systems (CRS) to a single coverage. Furthermore, the implementation of non-Byte data (e.g. 16 Bit) has been improved

and the possibility to include additional dimensions for filters such as cloud coverage has been implemented. Combining bands of different spatial resolutions is not possible yet, but work on this issue has started. Further improvements are done as part of the EVO-ODAS project [7].

Over the last months, the software has been continuously developed to accomplish the requirements on rendering and performance. However, all new features need to be thoroughly tested before they can be deployed.

2.2. Merging the datasets

Sentinel-2 products are a compilation of granules, which are ortho-images, 100x100km² in size [3]. Prior to rendering and serving, these images have to be combined to a mosaic. Due to the large number of products and their frequency, this needs to be performed on the fly. For one full coverage of the earth, 40.000 images with individual coordinate reference systems need to be combined.

The data can be either combined to an ImageMosaic, an extension for GeoServer, or to a Virtual Raster Table (VRT). A VRT is a GDAL raster file format, which is a simple XML file referencing multiple raster files and providing additional configuration options. While VRTs are more flexible, the rendering of ImageMosaics is much faster. With the current and upcoming improvements of the ImageMosaic plugin, there is no further need to use VRTs for the implementation.

2.3. Band combination

Sentinel-2 L1C data are divided in 13 bands of three spatial resolutions: 10m, 20m and 60m [8]. The bands can be added to the ImageMosaic and combined to different visual representations which are selectable by the user. To merge data of different spatial resolutions, the bands need to be harmonized by up- or downsampling. GeoServer is not capable of doing this yet. However, as stated earlier, work on this issue has already been started.

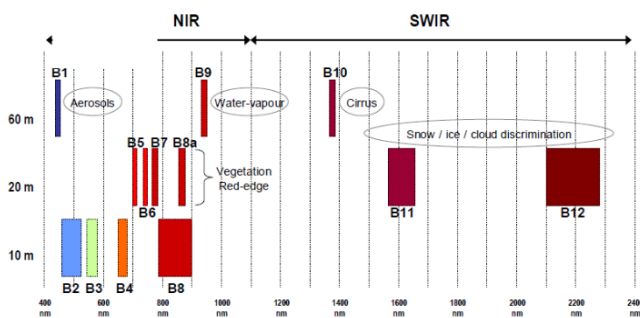


Fig. 1. MSI Spectral Bands versus Spatial Resolution [8].

2.4. Reprojection

The data are distributed in the 120 different UTM zones. While these coordinate reference systems (CRS) are ideal for regional studies, the data needs to be reprojected to a global CRS like WGS84 or (Web) Mercator for world-spanning tasks. As the final output will have a single global CRS only, the data needs to be reprojected to that CRS. It can either be reprojected on-the-fly while browsing or we can do the processing before adding them to the mosaic.

Reprojection of this amount of data needs a lot of processing power and alters the information due to pixel stretching, resampling and resizing. Thus, reprojection during preprocessing improves the view performance but increase the time needed for preprocessing and, at least when reprojecting to WGS84, will result in much greater storage needs because pixel width will be stretched by $1/\cos(\text{latitude})$.

2.5. Bit depth and value stretching

Despite the Sentinel-2 MSI instrument acquiring data in 4096 individual values (12 bit), the L1C data are provided in 16 bit format [8], enabling up to 65536 individual values per band. The values describe the Top of Atmosphere (TOA) reflectance as recorded by the satellite with original values ranging from 0 to 1. These values are multiplied with a fixed number. In the current Sentinel-2 L1C data format, the number is set to 10.000. Downloading and further processing requires these original values to be kept. But for a visual representation of the earth's surface, the original pixel values will result in an image that is too dark. Hence, the data needs to be cut and stretched. For Sentinel-2 L1C, cutting values above 4096 has shown to produce good brightness for visualizing the ground. However, very bright areas like clouds and glaciers may get oversaturated. Alternatively, non-linear stretching algorithms with less or no cutting of values can be provided. Stretching can happen during processing or on-the-fly.

Stretching down to 8 bit during processing enables us to use JPEG compression and to receive much smaller file sizes with the cost of losing flexibility to serve the original data, e.g. if we need to provide the original values through a WCS. To avoid border effects between granules, the same stretching algorithm needs to be applied to all granules.

GeoServer offers two methods to achieve that precondition: Linear stretching between a minimum and a maximum value and applying a gamma curve. For preserving flexibility we chose not to apply any cutting of values during processing. This task should be done on-the-fly in GeoServer. During processing, only linear stretching from TOA reflectency 0 to 1 is done if needed. Stretching in GeoServer is done using a style which is applied to the dataset. We can offer multiple styles for one dataset, resulting in a very high flexibility.

2.6. Transparency

As the individual granules overlap each other, the pixels outside the sensing area should be transparent. For Sentinel-2 L1C, these pixels have a specified value of 0. This value can be set as "nodata value" in the GeoTiff file. GeoServer renders the pixels matching this value as transparent. It would also be possible to store an alpha channel of the transparent pixels in the data. While these are good solutions for lossless compressed data, they are not suitable for lossy compressed data since they will introduce visible border effects between "nodata" and other values. To avoid such effects, GDAL implemented the possibility of adding a mask to a GeoTiff file. These are lossless compressed one bit raster overlays matching the areas which should be transparent. Additionally, the ImageMosaic extension has the option to add vector footprints into Sidecar Files as polygons in Well Known Text (WKT) format or as shapefiles. The footprints can easily be extracted from each granules metadata. But the provided coordinates are not precise enough to properly cut the raster data. To exactly match the pixels, the number of vertices of a Polygon would get huge, offering a considerably storage size and bad read performance.

2.7. Compression

The raster data can be compressed using several compression methods, which differ in compression ratio, read and write speed as well as being lossless or lossy.

Our own tests have shown that the lossless JPEG2000 compression of the original data offers the best compression ratio but the worst performance compared to other lossless algorithms. Alternative algorithms are Deflate, which offers the best compression ratio and LZW, which offers the fastest write speed. Compared to the lossless compression methods, lossy JPEG compression offers a much smaller file size and best viewing performance. However, the bit depth must be reduced to 8 bit prior to compressing, which results in a further loss of information.

3. EVALUATION

3.1. Compression

The performance of common lossless compression methods regarding write speed as well as their storage needs was measured.

Deflate and LZW algorithms are quite close competitors, while the compression ratio of PackBits is clearly below the others for this kind of data. Deflate offers better compression ratio while LZW offers better write speed. Differences are within a range of 10 to 20%.

The compression algorithms can store the differences between neighbouring pixels instead of the concrete values. This is called compression prediction. The setting on how

the predictor works has a big impact on the compression ratio of both Deflate and LZW algorithm, while the performance stays similar (Fig. 2). Thus, predictor setting 2 can always be preferred for this kind of data.

In the CODE-DE project, we chose to further consider the Deflate algorithm using predictor 2 for its better compression ratio.

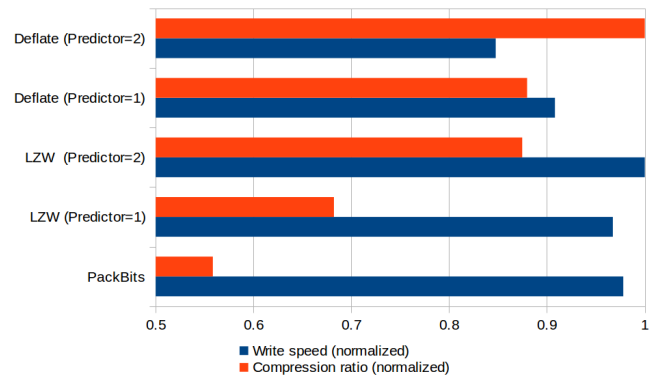


Fig. 2. Comparison of write speed of common lossless compression methods for GeoTiff.

3.2. Storage needs

While the compression ratio of different lossless compression algorithms differs, the biggest reduction in storage size can be achieved using a lossy compression. The JPEG compressed Sentinel-2 imagery has only approximately 10% the size of lossless compressed data (Table 1). This does not only mean smaller storage needs but has also an impact on access performance as less data needs to be read.

Downside of JPEG compressed data is the loss of information due to the reduction of the bit depth as well as the lossy compression resulting in compression artefacts.

Compr. Method	Size (MB)	Proc.Time (s)
JP2 Lossless	350	0*
Deflate	468	105
LZW	535	92
JPEG (Quality=75%)	50	108

Table 1. File Size and Proc. Time of different compression methods per Sentinel-2 L1C Granule.

*) processing time of JP2 format has not been evaluated as it is the native format of Sentinel-2 L1C data.

3.3. Access Performance

The reprojection of the granules to a single global CRS has the biggest impact on access performance. Performing this task during preprocessing will lead to a higher preprocessing

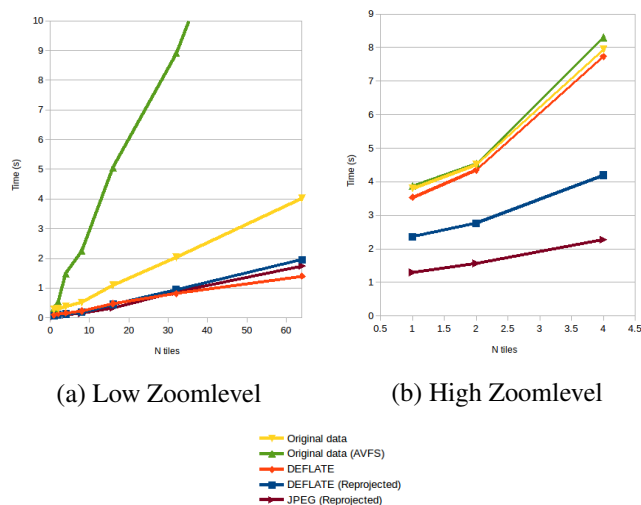


Fig. 3. Comparison of the WMS Performance of the selected compression methods and projections.

time, a loss of information and/or bigger file sizes. Reprojecting the data on-the-fly will lead to a lower serving performance. Fig. 3 shows the access speed of the data in two zoomlevels. In low zoomlevel, many granules are accessed at each request. The access time overweighs the processing time. Deflate, either reprojected or not as well as JPEG are within range of error. In high zoomlevels, only a few granules need to be accessed and these high resolution images need to be processed. The processing time overweighs the access time.

4. CONCLUSION

For every individual use case a suitable solution can be found, but no single solution will fit all requirements.

- For best quality and flexibility for further processing, it is best to preserve the original values either by keeping the original file format or by converting to a lossless GeoTiff format. The latter improves the performance at the cost of processing time and larger file size.
- Smallest file size can be reached when using a lossy compression like JPEG.
- Best performance can be reached with the data reprojected to the target CRS. Using lossy JPEG compression will improve the performance even more as less data needs to be read. Overviews and internal tiling are crucial for good performance independent of the chosen solution.
- Using two separate approaches for WMS and WCS may be considered.

In the CODE-DE Project we were approaching for a compromise between these contradictory terms. Main focus has been set on receiving the smallest file size with greatest flexibility that has good enough performance for serving. We decided serving JPEG compressed GeoTiffs to receive good performance and small file sizes, but in the original CRS to maintain some flexibility. Later in the project a separate solution for serving WCS could be realized using the original JP2000 files.

5. REFERENCES

- [1] CODE-DE, “Copernicus data and exploitation platform deutschland (code-de): code-de.org,” October 2017.
- [2] John W Hager, Larry L Fry, Sandra S Jacks, and David R Hill, “Datums, ellipsoids, grids, and grid reference systems,” Tech. Rep., DEFENSE MAPPING AGENCY HYDROGRAPHIC/TOPOGRAPHIC CENTER WASHINGTON DC, 1992.
- [3] SUHET, “Sentinel-2 user handbook,” Available online (accessed October 13, 2017) https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook, 2017.
- [4] Open Source Geospatial Foundation, “Gdal - geospatial data abstraction library website: www.gdal.org,” October 2017.
- [5] Open Source Geospatial Foundation, “Geoserver website: geoserver.org,” October 2017.
- [6] Jim Greenwood and A Whiteside, “Ogc web services common standard,” 2010.
- [7] Torsten Heinen, Bernhard Buckl, Simone Gianecchini, Stephan Kiemle, and Meißl Stephan, “Evolution of earth observation online data access,” in *Proceedings of 2016 conference on Big Data from Space (BiDS'16)*. Publications Office of the European Union, 2016, pp. 31–34.
- [8] A Gatti and C Naud, “Sentinel-2 products specification document,” Available online (accessed October 13, 2017) <https://sentinel.esa.int/documents/247904/685211/Sentinel-2-Product-Specifications-Documents>, 2017.

EFFICIENT PROTOCOLS TO STORE AND TRANSMIT FOR BIG DATA GENERATED BY EARTH OBSERVATION SATELLITES

Yousuke Ikehata, Takahiro Minami, Yuji Shimomura, Hidekazu Mikai, Naoyuki Fujita

Japan Aerospace Exploration Agency

ABSTRACT

The recent earth observation satellites generate big size data, according to enlargement of sensor resolution and diversification of sensor types and analysis. The trend is getting more remarkable day by day and data sizes of latest satellites under development come up to tens of GBs and more.

Conventional systems are unable to treat such a big data smoothly and hence we have to reconsider the systems from the elements levels like storage, processing and transmission for sustainable earth observation system,

In this paper, we show the result of data transfer performance for three transmission protocols and our future data center basic design.

Index Terms— transmission protocol, LFN, data center

1. STORAGE SYSTEM VOLUME

At first, big data effects the size of storage system volume directly. Some satellites have two or more sensors and the sensors produce multi spectrum data, SAR data, hyper spectrum data, etc. In other words, one satellite generates not only one data set but also two or more data sets. Original data set are called “RAW data” and the “RAW data” generates some “value added data”. Furthermore, observation data are generated in each satellite path (in every 90 minutes).

Generally, earth observation data are used for monitoring targeting global climate changes like an El Nino, arctic sea ice decline and so on. The monitoring need over 30 years’ earth observation data and hence those data have to be preserved even after the satellite missions finished.

From the above characteristics, huge data are generated and piled up in a storage every day, but can’t be deleted permanently.

2. DATA PROCESSING

Secondly, big size data effect processing. Normally, “RAW data” is data which are outputted from sensor and we convert and correct it to “Level 1 data”. We process “Level 1 data” to physical data called “Level 2 data”. “Level 2 data” is single orbit data, and global data called “Level 3 data” is generated by collecting and overlapping “Level 2 data”.

Those conversion and processing, which handle earth observation data, effects read/write speed from/to disks and memories. Of course, those processes run in many servers and

they use shared file system, in order to speed-up conversion and processing. For processing earth observation data need high throughput and low latency shared file systems.

3. DATA TRANSMISSION

In addition, transmission is affected by data size. Processed data are stored a few years, but those are occasionally replaced with a new version when the algorithms for processing are updated (“version up”). A version up usually targets a series of data observed and stored for several years and the size of those might be in the order of tens of TB. That replacement targets all observed data, and replacement must finish in some days or weeks because current observing data version matches replaced version. That replacement targets all observed data, and replacement must finish in some days or weeks because the version of algorithm which is applied to re-process should be the same as current observing one. In order to achieve this, we employ a supercomputer called JSS (JAXA Supercomputer System) in JAXA. [1] However, JSS is placed in a JAXA center located in Tokyo/Chofu, while the earth observation data themselves are stored in another JAXA center located in Ibaraki/Tsukuba. We accordingly have to transmit the data over WAN networks before and after a version up. The above centers are around 50km apart and the RTT (Round Trip Time) is about 5.8 [ms]. In case that TCP/IP is used as a transfer protocol, the RTT above constrained the transmitting speed under tens of MB/second. If we complete a quick version up in a few days, we must improve not only processing but also transmitting in a version up process.

4. FEASIBILITY STUDY

We adopt distributed file system distributed file system protocol for resolving above 2 problems: large scale storage (3.5 PB and 3PB) and high performance shared file system (300MB/second).

Two file systems are constructed, the size of which are 3.5PB and 3PB respectively. Since they are based on a distributed file system, read/write access is balanced automatically. We are currently responsible for operations on 3 satellites (GOSAT, GCOM-W, ALOS-2) and 1 sensor (DPR on NASA GPM) and share a same storage to decrease spare space and reduce operating cost. The throughput scores

over 300 MB/second from each client system and they haven't caused delay and congestions.

Last year, we evaluated three transmission protocols using large scale data on LFN (Long Fat Network).

1. TCP base protocol (protocol A): an open source software using TCP protocol.
2. UDP base protocol 1 (protocol B): a commercial product implemented custom protocols over UDP.
3. UDP base protocol 2 (protocol C): a commercial product implemented custom protocols over UDP.
4. home/caching technology (protocol D): subsets of distributed filesystem which is a commercial product.

Those protocols improved transmission performance between two centers mentioned above under operational environment.

5. DATA AND OPERATIONS

The word of "Big Data" has many meanings. For example, the volume of single data is larger than hundreds of GBs or TBs. Or each data size are a few MBs or GBs but the number of data are more than trillions.

In such a different situation, solutions are different too. Small data finishes transmitting in short time but transmission speed cannot get fast. Vice versa, large data needs much time to finish transmitting but speed can get fast.

In our situation, JAXA/SAOC has many data and each data size isn't large. JAXA/SAOC launched earth observation satellite MOS-1(momo) in 1987 and have launched 12 earth observation satellites till now. So SAOC are keeping archive data over 30 years (Total size of data are 3.5 PBs.). So, transmission speed doesn't get problem in nominal operations.

However, we sometime transmit all data (long term data) for reprocessing. Reprocessing performs in JSS and finishes in short time. So, executions time can be shortened but transmission time is slow. Furthermore, data is transmit twice; send to JSS before reprocessing and receive from JSS after reprocessing.

This paper evaluates in small data size situations and conditions are shown as below.

- Evaluates 2 data pattern;
 - 100MiB x 100 (Total 10000MiB/10GiB)
 - 2000MiB x 5 (Total 10000MiB/10GiB)
- Transmit between datacenters (RTT=5.819msec)
- Test runs 10 times
- Remove best/worst 2 results.
- Calculate 6 data mean. (Evaluate value)

6. ENVIRONMENT

Servers used in feasibility study are shown in Table1 and Table2.

Table 1 Server specification(Tsukuba)

object	Value
Hardware	Dell PowerEdge R630
CPU	Intel Xeon E5-2630 v4
OS	Ubuntu Server 14.04.5
Memory	64GB
HDD	SAS 3TB(RAID10)
NIC	10GbE
NAS	NFS/10GbE

Table 2 Server specification (Chofu)

object	Value
Hardware	FUJITSU PRIMERGY RX200 S8
CPU	Intel Xeon E5-2643 v2
OS	RHEL Server release 6.5
Memory	32GB
HDD	SAS 900GB(RAID1)
NIC	10GbE
NAS	NFS/10GbE

Network configuration is shown in Table3.

Table 3 Network configuration

object	Value
Bandwidth	10GbE(best effort)
RTT	5.819msec
iPerf value	9.41 Gbits/sec
Connection	L3 connection
WAN	Dedicated line(L2VPN/SINET[2])
Number of FWs.	2

7. TCP BASE PROTOCOL (PROTOCOL A)

Protocol A is an open source software using TCP protocol and connects by multi sessions to improve transmission speed. It is extension of FTP protocol and subset of grid computing framework.

Some features are shown in as below.

- Secure connection; SSH base or PKI base authentication.
- Parallel transmission; transmit multiple files in parallel.
- Multi session; transmit each files using multi sessions.

In this feasibility test, authentication method using SSH.

At first, we evaluate effects of multi session. The result is shown in Table4.

Table 4 Transmit speed (Protocol A/multi session)

Data	Number of session	Speed
100MiB x 100	1	223 MiB/sec
	4	483 MiB/sec
2000MiB x 5	1	430 MiB/sec
	4	482 MiB/sec

The number of sessions improves speeds.

Secondly, we evaluate effects of parallel transmission. The result is shown in Table5.

Table 5 Transmit speed (Protocol A/parallel)

Data	Number of parallel	Speed
100MiB x 100	1	483 MiB/sec
	2	648 MiB/sec
	4	653 MiB/sec
2000MiB x 5	1	482 MiB/sec
	2	685 MiB/sec
	4	557 MiB/sec

The number of sessions are 4 and we change parallelism only. The number of parallelism improves speeds.

8. UDP BASE PROTOCOL 1 (PROTOCOL B)

Protocol B is a commercial product implemented custom protocols over UDP.

It provides encrypt function and speed limitation function. In this feasibility study, we evaluate no encryption and no speed limitation (transmit speed is limited less than 1Gbps because of license.).

The result is shown in Table 6.

Table 6 Transmit speed (Protocol B)

Data	Speed
100MiB x 100	63.1MiB/sec
2000MiB x 5	63.4MiB/sec

The result is different from Protocol A. There is almost no difference depending on the test data (the number of data and the size of each data).

9. UDP BASE PROTOCOL 2 (PROTOCOL C)

Protocol C is also a commercial product implemented custom protocols over UDP like protocol B.

It is using UDP for transport layer protocol and special protocol manages higher layers for improving reliability and efficiency.

It has 2 options; encryption and multi thread transportation. In this feasibility test, we change only encryptions and thread is fixed single thread because of our current network system designs.

The result is shown in Table 7.

Data	Encryption	Speed
100MiB x 100	Encrypt	80.90MiB/sec
	No encrypt	81.87MiB/sec
2000MiB x 5	Encrypt	89.71MiB/sec
	No encrypt	91.35MiB/sec

The result is almost same as Protocol B. There isn't difference depending on the test data (the number of data and the size of each data) and encryption degrades speeds a little.

10. HOME/CASHING TECHNOLOGY (PROTOCOL D)

Originally, protocol D isn't transmission software like ftp/sftp. It is sub function of distributed file system to share data between different data centers. JAXA already uses distributed file systems and want to share data between Chofu and Tsukuba so this protocol suits for our needs.

This protocol using caching technology for sharing different data centers and there are 4 options by caching functions.

In this feasibility test, "home site" is source data center (Tsukuba) and "cache site" is destination data center (Chofu). The operations are described in below.

1. Put/write data to "home site" storage (Data don't transmit to "cache site" in this time and we can't see anything in "cache site" storage.).
2. Replicate metadata between "home site" and "cache site" automatically (We can see data in "cache site" storage but data don't exist.).
3. Get/read data from "cache site" storage (Data send to "cache site" storage from "home site" storage.).

The result is shown in Table 8.

Table 8 Transmit speed (Protocol D)

Data	Speed
100MiB x 100	73.5MiB/sec
2000MiB x 5	87.0MiB/sec

Large data is faster than small data for transmission like typical protocols. Furthermore, speed is almost same as other UDP based protocols.

11. RESULTS AND DISCUSSIONS

As we described above, Earth observation satellite data are getting bigger and bigger day by day and the storage cost is accordingly getting higher and higher even for ground systems. Not only JAXA but also other space agencies are expected to be facing the same problems and we all have to find a more efficient way to storage such a big data. JAXA has started to use distributed file system for the storage protocol.

As a next step, we've started to consider mutual use of computing and storage resources placed in our centers more seamlessly. In addition, we are also investigating use of cloud services for storing off-load resources to reduce TCO (see Figure 1) through making a trade-off between cost and performance.

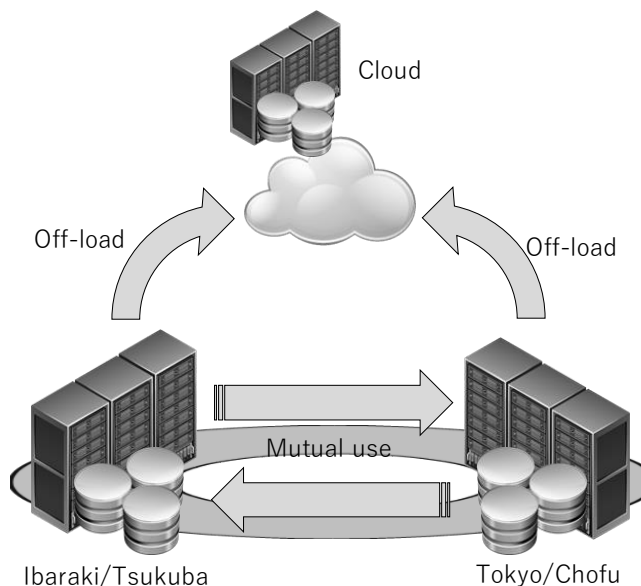


Figure 1 Conceptual relations between two data centers and cloud

The mutual use of those needs further solutions which can transmit TB scale data like satellite observation data and Virtual Machine images.

12. REFERENCES

- [1] Overview of JSS2, https://www.jss.jaxa.jp/overview_of_jss2_e/
- [2] L2VPN/SINET https://www.sinet.ad.jp/en/tag/l2vpn_en

STORAGE OPTIMIZATION ACCORDING TO MISSION-ORIENTED CRITERIA

Olivier Queyrut⁽¹⁾, Xavier Geoffret⁽²⁾, Pierre-Marie Brunet⁽¹⁾,
Patrick Ginet⁽¹⁾, Cyrille Parra⁽²⁾, Denis Gutfreund⁽²⁾

(1) CNES, (2) Atos

ABSTRACT

The new space missions in the domain of Earth Observation generate a volume of data that is significantly larger than in previous years. For instance, the French-US SWOT mission [1] (NASA-CNES cooperation) will generate 1TB of telemetry daily. One of the main complexities for designing the ground segment is to define the proper approach for the data storage to support the processing, bulk reprocessing and distribution activities.

In 2016-2017, CNES held a R&T named "Storage optimization according to mission-oriented criteria" whose goal is to investigate if the knowledge of how the data are handled and processed by the ground segment can define more accurate tuning parameters of the different storage infrastructures that are today optimized according to technical and / or general usage scenario.

This presentation is intended to provide feedbacks of this R&T that was conducted with Atos. During this R&T, the first phase was to achieve a state of the art of storage technologies for scientific data; then, in a second phase, to collect the needs of storage and associated processing for some significant missions; to deduce, therefrom, some ways of optimizing the storage to meet these missions' needs. The R&T is also continuing on the implementation of POCs, whose role is to provide clarification and validation on the contribution of some technologies that are part of the presented optimizations.

Index Terms — Storage, scientific data, HPC, HTC, Big Data, optimization, space missions, CNES

1. INTRODUCTION

With their increasingly large data volumes, new space missions in the domain of Earth Observation are entering the era of Big Data. The architecture of the ground segment faces the challenge of managing a very high volume of data combined with complex processing requiring an infrastructure with hundreds of computing cores. One of the difficulties lies in the approach to be taken for storing data in relation to processing, reprocessing and distribution requirements throughout the mission lifetime.

To illustrate, the French-US SWOT mission (NASA-CNES cooperation) that will be launched in 2021, will

generate daily 1TB of raw telemetry and produce 13TB of intermediate and user products.

To cope with the data volume to be stored, two approaches which can be combined, are generally implemented: on the one hand the file systems distributed on a storage cluster such as NAS, SAN, parallel file systems or distributed file systems (such as HDFS); on the other hand, the hierarchical storage composed of several storage media having different properties in terms of throughput, latency and cost.

All these technologies involve data transfers: transfer to a processing node, data grouping in order to combine them, data staging in a hierarchical storage. It is thus very clear that a major factor to optimize the data storage (in particular in terms of latency and network traffic) lies in the intelligent location of the data to support both the processing and the distribution activities.

The R&T named "Storage optimization according to mission-oriented criteria" aimed at investigating storage solutions adapted to the context of future Earth Observation missions. Its intent is not to invent a new storage system but to first select the most suitable technologies, and secondly to find their optimization levers, or tune their functioning according to the specificities of a mission ground segment.

The next sections describe the activities that were conducted and provide feedbacks of this R&T.

2. STATE OF THE ART OF SCIENTIFIC DATA STORAGE

We focused first on use cases, to bring out a classification of functional needs. Which data is stored where, for which functional requirements? It also allows to provide a first overview of solutions and commonly encountered players.

The state of the art was focused on 5 technological themes allowing to browse the whole storage hierarchy – namely NVDIMMs, flash storage, rotating disks, tapes and digital optical disk–. Several high-level solutions were studied because they were relevant in the context of storing and processing CNES mission data.

Persistent in-memory file systems are a way to use the performance of volatile memory as a persistent storage device, with traditional file access paradigms. Mature implementations, like Alluxio [2], may reduce processing times, especially for Hadoop/Spark data processing.

Scalable storage (which covers the scale-out NAS solutions, object storage and cloud storage) address the increasing needs for high scalability, efficient data protection, metadata management, and reduced hardware costs. Major players include (by alphabetical order of vendors) DDN Storage [3], Dell EMC [4], IBM [5], RedHat [6], and Scality [7].

Multi-tier storage solutions manage multiple levels of storage through different technologies (SSD, fast disks, capacitive disks, tapes, cloud). The expected benefits are the optimization of costs (strategic alignment of investments with the business value of the data throughout its life cycle), efficient data protection (in the meaning of protecting against data loss) and long-term conservation. CNES is currently using two multi-tier storage solutions: HPSS [8] and Oracle HSM [9].

Many data frameworks are relevant in the context of storing and sharing data for space missions, particularly concerning data management, federation of storage solutions, as well as synchronization and data sharing solutions. Most interesting data frameworks include Xrootd [10], Globus [11] and iRODS [12].

The use of different types of processing technologies for High Performance Computing (HPC) and Data Analytics (DA) has led us to consider recent advances in HPC-DA converged architecture: shared storage, common resource manager, unified high-performance network.

3. THE STARTING POINT: THE MISSIONS

For the data managed during a mission, we have distinguished two phases.

First, a production phase during which the various products are generated. In this phase, two distinct needs can be distinguished: (1) processing requirements that follow the workflow of the incoming data where the rhythm of processing is clocked by the periodic acquisition of the telemetry (the processing times are thus relatively predictable), and processing capacities must be sufficient to process data flows in real time; (2) bulk reprocessing requirements to generate new versions of the products based on more accurate algorithms where processing capacities must allow these reprocessing operations to be completed as quickly as possible. The processing implemented in this phase were historically of HPC type (explicit parallel programming with MPI libraries, parallel file system like Spectrum Scale [13]), but for that, increasingly, data analytics processing (implicit parallel programming with MapReduce or Spark frameworks, distributed file system like HDFS) are deployed.

Secondly, a valorization phase during which the products are made available to users (experts, researchers, partners, general public). Use cases and access patterns are not known in advance and can be very variable: steady access to the

products (e.g. systematic downloads every time a new product is being made available), peak of interest on products covering a geographical area where an event such as a climate disaster has just occurred, etc. We see that in this valorization phase, CNES may require not only to publish data but also to make platforms available to develop new processing chains.

According to the phases, the constraints and storage needs are different. They are currently provided, for each mission, by a specific infrastructure. The rest of the document presents storage optimization tracks which may vary according to the phases.

4. OPTIMIZATION TRACKS

4.1. Optimization of existing infrastructure

In this chapter, several approaches to optimize existing infrastructure were studied.

Some are purely at a technological level and can be considered somehow as generic solution to improve storage systems, in the sense that they are not specifically designed to address the domain of ground segments. Bull Director for HPSS [14] dynamically sorts the requests of data located on tapes so that the access latency is minimized. Its usage is relevant in a scenario where “old” data, thus located on a cold storage tier, are being accessed for bulk reprocessing purposes or for distribution to end users. The Burst Buffer technology intends to accelerate the parallel filesystem access (like Spectrum Scale) by adding high rate storage devices (SSD disks) between the processing nodes and the parallel filesystem. The added value of this technology needs more assessment since the actual improvement depends on the access types (file size, i/o implementation). The last technology that was studied is NVMe over Fabrics. Its goal is to accelerate the data processing by temporary providing a flash storage capacity over the network as if it were local.

Other approaches intend to act at a different level. Storage systems manage their data according to general criteria such as the last time a data was accessed. These rules are only based on the present situation and do not take into account what could happen in the future. By adding business rules about how data will be handled by the ground segment, the storage system can better anticipate the localization of data so that they are available on the most performant device as soon as the ground segment uses them: no waiting time due to data being moved from one device to the other (e.g. from tapes to high performance disks on a hierarchical storage system, from local storage to external cloud storage in case of cloud computing). For instance, a business rule can state which data is necessary and when for a scheduled bulk reprocessing campaign. With the knowledge of the future data usage, the storage system can

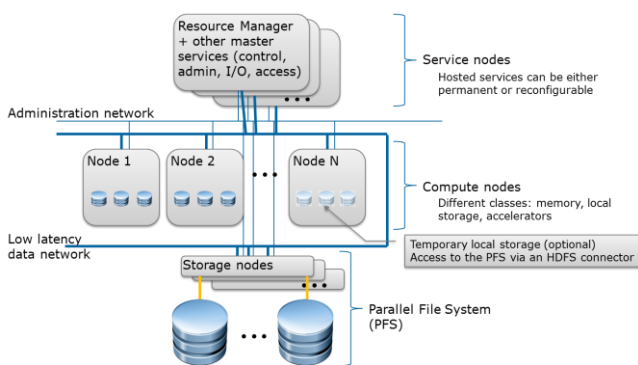
better schedule the data staging according to its working load, the estimation of the transfer duration, etc.

To characterize the benefits of these approaches and validate their impacts, we implemented a Proof of Concepts (POC) dedicated to optimizing multi-tier storage (e.g. flash storage and rotating disks, or disks and tapes, or local and cloud storage). This POC implements both the Bull Director for HPSS to optimize data access from tapes in a very demanding context, and an automat that intelligently moved the data according to business rules. For the automat, two approaches are developed: a static one and a dynamic one.

The static approach is to develop a finite state machine that describes the life cycle of the data, and the integration of such a machine to drive the optimal placement of the data among the different storage tiers.

The dynamic approach is based on predictive analytics with machine learning technologies. Such technologies should help to predict when a specific data set is about to be used, and thus anticipate an optimal placement on storage tiers. This prediction can be based on the analysis of users' behaviors and/or on some external sources of information, such as meteorological alerts, to infer that satellite observations over an area will very likely be accessed. To illustrate in the frame of the SWOT hydrology mission, we can imagine that in the case of a flood in an area, users will not only download the most recent SWOT data over the area, but also the data in the past, data covering neighboring areas and data of other missions –e.g. Sentinel-2 images [15]– to better understand the phenomenon. Pre-staging these data on a low latency storage device before they are requested will surely improve their download rate.

4.2. Design of new architectures



Several themes emerged during the analysis phase. First of all, a few optimizations tracks were studied around the convergence or the cohabitation of HPC on one hand, and HTC or Big Data on the other hand. In terms of infrastructure, the use of a parallel filesystem is an efficient way of mutualizing the same hardware resources for the two programming paradigms. Other challenges include the use

of a job scheduler like slurm [16] or PBS Pro [17] to drive a Hadoop resource manager like YARN [18], or a dynamic reconfiguration of compute nodes software stacks to easily and rapidly change computing environment.

The openings to cloud storage and data governance are another hot topic for ground segments of Earth Observation missions. More specifically, Cloud Storage Gateways could be used to share cloud-hosted data between agencies, while providing fast local access. There is consensus that this approach requires some changes in the way data is accessed: object-style API, like AWS S3 [19] or OpenStack Swift [20], allows a unified access method, whether data is stored locally, in a private cloud or in a public cloud. However, the processing algorithms must be adapted to this new data access (which requires to read the entire file and work in memory). The cutting of data in smaller pieces is often required and may completely change the algorithmic logic.

4.3. Object storage

Finally, we studied the contribution of object storage technologies for some identified use cases.

5.2.1. Overflow of a parallel file system

During the production phase, the use of a parallel file system is the key to efficient and performant data access, but at premium costs. Object storage technology can be combined with ILM (Information Lifecycle Management) policies to provide virtually infinite parallel file system capacity at lower marginal costs.

5.2.2. HDFS replacement for Hadoop/Spark data analytics processing

Object-based storage solutions to replace HDFS have many benefits. Thanks to the integrated erasure coding mechanisms built into object storage, it is often possible to reduce storage costs by a factor of 2. It is possible to make storage evolve independently of computing resources. The data no longer needs to be copied to and from HDFS. Data analysis clusters have direct access to the data, and thus reducing processing times.

5.2.3. Backend storage for enterprise file sync and share

This use case allows CNES employees and partners to synchronize and share documents and data between multiple devices (mobile, PC and servers) in a public or a private cloud. In a context of cooperation between space agencies, this kind of solution can advantageously replace the traditional data transfers by SFTP, FTP or other protocols. ownCloud [21], as an open-source implementation, has been detailed.

6. CONCLUSION

Designing a storage system for a Big Data system needs to juggle with different optimization criteria. Among them, the latency, volume and cost are the most relevant. With the huge amount of data, it is no more relevant to have one storage system per usage scenario because data movements are too expensive and storage volumes will explode.

The R&T held by CNES with Atos identified several technologies that either optimize existing infrastructure or help designing multi-purposes storage systems that address different usage scenario. It also showed that anticipating the data usage (by providing static information about the future usage or by dynamically analyzing users' behaviors) can help storage systems to better answer to users' requests.

7. REFERENCES

- [1] "SWOT: a promising hydrology and oceanography mission." [Online]. Available: <https://swot.cnes.fr/en/SWOT/index.htm>. [Accessed: 09-Oct-2017].
- [2] "Alluxio - Open Source Memory Speed Virtual Distributed Storage," *Alluxio*. [Online]. Available: <http://www.alluxio.org>. [Accessed: 09-Oct-2017].
- [3] "DDN WOS: Object storage," *DDN.com*. .
- [4] "Elastic Cloud Storage - Object Storage Solutions." [Online]. Available: <https://www.dellemc.com/en-us/storage/ecs/index.htm>. [Accessed: 09-Oct-2017].
- [5] "IBM Cloud Object Storage," 24-Oct-2016. [Online]. Available: <https://www.ibm.com/cloud-computing/products/storage/object-storage/>. [Accessed: 09-Oct-2017].
- [6] "Red Hat Ceph Storage." [Online]. Available: <https://www.redhat.com/en/technologies/storage/ceph>. [Accessed: 09-Oct-2017].
- [7] "Scality RING: Scale-out File and Object Storage," *Scality*. [Online]. Available: <http://www.scality.com/products/ring/>. [Accessed: 09-Oct-2017].
- [8] "HPSS - High Performance Storage Systems." [Online]. Available: <http://www.hpss-collaboration.org/>. [Accessed: 09-Oct-2017].
- [9] "Oracle Hierarchical Storage Manager." [Online]. Available: <https://www.oracle.com/storage/tape-storage/hierarchical-storage-manager/index.html>. [Accessed: 09-Oct-2017].
- [10] "Home Page | XRootD." [Online]. Available: <http://xrootd.org/>. [Accessed: 09-Oct-2017].
- [11] "Research data management simplified. | globus." [Online]. Available: </research-data-management-simplified>. [Accessed: 09-Oct-2017].
- [12] "iRODS." [Online]. Available: <https://irods.org/>. [Accessed: 09-Oct-2017].
- [13] "IBM Spectrum Scale," 09-Oct-2017. [Online]. Available: <https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage>. [Accessed: 09-Oct-2017].
- [14] "Extreme Data," *Atos*, 02-Jul-2017. [Online]. Available: <https://atos.net/en/products/high-performance-computing-hpc/extreme-data>. [Accessed: 12-Oct-2017].
- [15] "Sentinel-2: Viewing Earth in unprecedented radiometric detail." [Online]. Available: <https://sentinel2.cnes.fr/en/sentinel-2-0>. [Accessed: 12-Oct-2017].
- [16] "Slurm Workload Manager." [Online]. Available: <https://slurm.schedmd.com/>. [Accessed: 12-Oct-2017].
- [17] "PBS Professional: HPC Workload Manager and Job Scheduler." [Online]. Available: <http://www.pbsworks.com/PBSProduct.aspx?n=PBS-Professional&c=Overview-and-Capabilities>. [Accessed: 12-Oct-2017].
- [18] "Apache Hadoop 3.0.0-beta1 - Apache Hadoop YARN." [Online]. Available: <https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html>. [Accessed: 12-Oct-2017].
- [19] "Amazon Simple Storage Service (S3) — Cloud Storage — AWS," *Amazon Web Services, Inc.* [Online]. Available: <https://aws.amazon.com/s3/>. [Accessed: 12-Oct-2017].
- [20] "OpenStack Docs: Welcome to Swift's documentation!" [Online]. Available: <https://docs.openstack.org/swift/latest/>. [Accessed: 12-Oct-2017].
- [21] "Home," *ownCloud GmbH*. [Online]. Available: <https://owncloud.com/>. [Accessed: 09-Oct-2017].

EUMETSAT SUBMISSION INFORMATION PACKAGE (SIP)

David Berry and Michael Schick

EUMETSAT, Eumetsat-Allee 1, 64295 Darmstadt, Germany

ABSTRACT

EUMETSAT operates a long term archive based on the principles of the CCSDS Reference Model for an Open Archival Information System (OAIS), containing meteorological data spanning nearly four decades, covering data from many different missions and satellites. In the past, mission data has been received in mission specific formats (file based), with the metadata being extracted primarily from the filenames, but also from inside the data files. This added complexity to the data management process. With this approach it is necessary for format specifics to be known at ingestion time. For each new mission, application code had to be developed and maintained to extract metadata from each new format and respective version. In recent years there have been efforts to reduce the complexity of managing data, standardizing interfaces/ formats, following long term preservation principles, thereby improving interoperability and distribution of data. In this short abstract we explore the introduction of the EUMETSAT SIP format, and the difficulties that we have encountered.

Index Terms— SIP, Long Term Data Preservation, OAIS

1. INTRODUCTION

The EUMETSAT Submission Information Package and its metadata representation is based on an extension of the OGC Earth Observation Metadata profile of Observations & Measurements [2]. The SIP generic metadata format is intended to be used by future EUMETSAT missions, thus providing an abstraction from mission specific format knowledge in the context of metadata management. In this extended abstract the motivation and obstacles for the development and introduction of the EUMETSAT SIP is reflected. We then outline the planned usage of the EUMETSAT SIP, the challenges that this has entailed, and the benefits to EUMETSAT and its users.

The EUMETSAT SIP is a submission information package as defined in [1], in our implementation the SIP is a UNIX tape archive file (although this is flexible) containing the following elements;

- A manifest file (manifest.xml) containing information about the content of the SIP and the relationships among the different items constituting the package.
- An earth observation metadata (EOPMetadata.xml) file containing metadata about the product that extends OGC EOM [2]; EUMETSAT specific metadata

attributes are conformant to the OGC 10-157r4 principles.

- Data Objects, which containing the actual product information: there can be many of these per SIP e.g. Data, Browse, and Geolocation.

A valid SIP must contain only one **manifest.xml**, one **EOPMetadata.xml** and one or more data files. Figure 1 depicts this graphically.

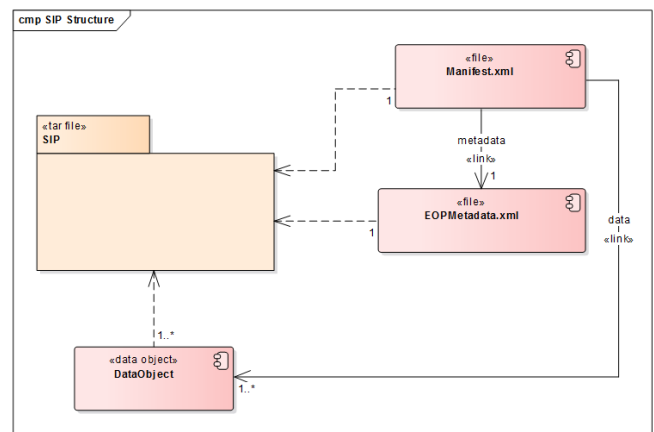


FIGURE 1: SIP STRUCTURE

The EUMETSAT SIP is our implementation of a SIP as defined by the OAIS model. For our purposes a SIP can either map directly to one AIP (Archival Information Package), or many SIPs can map to a single AIP; this shall be done for MTG (Meteosat Third Generation) products for instance. The DIP (Dissemination Information Package) is usually the same as the AIP.

The purpose of the SIP is to encapsulate the EO product and present all relevant metadata from the product data files in a well-defined, clearly specified format.

2. MOTIVATION

The need for a standard metadata format for EO data is clear. The amount of EO data that is being generated and preserved is growing at an incredible rate (Sentinels, MTG...). As a result, it will become ever more important to have a standard format of EO metadata. This will both, support the interoperability and sharing of catalogue data between organizations, as well as allowing users to easily work with and interpret vast amounts of data and products, simplifying

their processing chain. It is our ambition that the EUMETSAT SIP will help to meet this need. An additional benefit that the EUMETSAT SIP will bring is a simplification in the number of software components that are required for the long term data preservation.

The Unified Meteorological Archive and Retrieval Facility (UMARF) is the multi-mission element (MME) at EUMETSAT responsible for the long term data preservation of all EUMETSAT mission data. In the past, each new programme has introduced a new format for products. This has necessitated new software to be developed for the extraction of metadata depending on the product format, in the past metadata has typically been extracted from a combination of file name and metadata in the product. These so called, front ends, are specific to each mission, for instance there is a front end for Meteosat First Generation (MSG) and if there is a format change in one of the MSG products then the MSG front end application must be updated.

At the facility level, the introduction of the EUMETSAT SIP generates a huge advantage. The front end for the EUMETSAT SIP doesn't have to 'understand' the product format to extract cataloguing metadata since all of the relevant information is contained in the EOPMetadata. Hence, the UMARF will be unaffected by any change in product format. It is our intention to use the EUMETSAT SIP as the format for all future missions, and as a result of this it will be possible to utilize the generic SIP front end for the management of new mission data.

2.1. Comparison of EUMETSAT SIP & Sentinel-SAFE

EUMETSAT is responsible for the long term data preservation and distribution of Sentinel-3 data, which is also provided as a submission information package in the Sentinel SAFE format [4]. The SAFE format has some similarities to the EUMETSAT SIP, for example both extend OGC O&M [2], however, there are some differences, for instance, SAFE is XFDU [3] compliant, whereas EUMETSAT SIP is not.

The structure of the Sentinel-3 SAFE package is also very different, since there is one xml file that is a combination of a manifest (detailing which files are included in data) and metadata. The comparison is summarised in Table 1.

Table 1. EUMETSAT SIP vs Sentinel-SAFE Format

Format	OGC O&M	XFDU	SIP Content	
			Metadata	Manifest
SIP	YES	NO	YES	YES
SAFE	YES	YES	NO	YES

Despite these difference it is, in theory, relatively simple to convert from the Sentinel-3 SAFE format into the

EUMETSAT SIP format. There are ongoing initiatives at EUMETSAT exploring how our products can be provided in a range of formats that suit user needs. Using a standardised format for all our products we can simplify this process considerably and enable the data to be shared in useful formats much more easily.

3. INTRODUCTION OF THE EUMETSAT SIP

Imposing the EUMETSAT SIP on future programs has not been easy. Initially, it was our aim to have the data producers in the ground segment responsible for creating the products in the EUMETSAT SIP format. Although the need and advantages of the SIP have been clearly identified and are generally well understood, there has been a degree of resistance from the new programmes regarding the EUMETSAT SIP metadata format, leading to many internal discussions; and a number of arguments in opposition to the SIP format were raised. In the following several of these are listed.

Argument 1: Products are based on standards e.g. NetCDF which has metadata fields; why should these be copied?

Response: Although the plan is to have products in one standard format (NetCDF) today, this could change in the future – if and when new formats become the accepted industry standard.

Argument 2: The SIP format enforces that metadata has to be duplicated – a copy in NetCDF and a copy in EOPMetadata.xml – which needs to be kept synchronized, therefore that the data consumers should be responsible for the generation of the SIP;

Response: If consumers of the data have to read metadata from the product changes will be required when the changes to the format are introduced. Having consumers of the data being responsible for the generation of the SIP means that the process has to be repeated by each consumer of the data.

Argument 3: The SIP format will lead to an increase in the implementation time because of the need to build the SIP. A particular concern was raised about the complexity of the OGC O&M schemas, and the incurred costs for the implementation to write a SIP EOPMetadata file

Response: It was demonstrated how easy it is to generate a schema compliant XML using simple technology such as XSLT. A simple conversion example was delivered to prove that the complexity and therefore cost of the introduction of the SIP *should* be minimal.

Argument 4: Generating the SIP requires the harmonisation of the metadata for different products. Each of the satellite

instruments collect different measurements and therefore generate different metadata;

Response: Help was also given to the program in defining which metadata is actually required. We worked closely to assess what data will be useful for the users and how this could be used to generate valid EOPMetadata.xml files. With a XSLT example we demonstrated that it is relatively easy to generate valid SIP EOPMetadata file for products.

Despite our best efforts to have the SIP generated in the ground segment, i.e. by the data producers, it was eventually decided that the SIP will be generated by the UMARF MME at EUMETSAT.

An agreement was made with MTG for products to be sent in an intermediate format, that we call MTG-SIP. This format contains the product files and a manifest containing a limited amount of information. In addition to this an Interface Control Document (ICD), documenting the mapping of the NetCDF parameters to EOPMetadata parameters will be developed. The specification of the MTG-SIP is still in development. For EPS-SG (EUMETSAT Polar System – Second Generation) it was agreed for plain NetCDF files to be sent, i.e. not in an archive file.

The SIP Builder application will be responsible for constructing a EUMETSAT SIP from an MTG SIP and EPS-SG products. This is a new software component that is intended to be deployed as the single interface to MTG and EPS-SG.

4. CURRENT SITUATION

In our discussions with MTG and EPSSG we were not able to persuade the ground segments (data producers) to distribute products in the SIP format. However, the SIP format will still be used, with the production of SIPs being handled by a new component. Whilst this is not ideal, a reasonable compromise has been made, allowing products to be processed by the UMARF facility in the EUMETSAT SIP format.

The UMARF team have begun to implemented EUMETSAT SIP front end software that will be responsible for the metadata management of products that arrive in the SIP format. The objective of this software is to be the single entry point used to manage the metadata for all future programs. This will mean that new products can be quickly and easily catalogued and preserved, with no new, complex frontends having to be implemented to interpret the product format. This will greatly simplify the future developments, saving time and money. Of course the additional overhead of maintaining the SIP Builder will be present, however all formatting will be dealt with in a single component, therefore

the overall maintainability of the UMARF facility will be greatly improved.

5. SUMMARY AND OUTLOOK

In this short abstract we have described the EUMETSAT SIP, a new format extending the OGC O&M [2]. This format is intended to be used for long term data preservation at EUMETSAT for all future programs, and is also foreseen to be used for reprocessed products. We have outlined some of the issues that were encountered during the attempts to have the EUMETSAT SIP format adopted by the future programmes, MTG and EPS-SG. Ideally the SIP would have been created by the data producers (providers) so the archive facility doesn't have to 'know' anything about the product format. A compromise was eventually reached with MTG whereby an intermediate SIP format containing a much simplified manifest will be provided. For EPS-SG the products will be sent as a single NetCDF file. The SIP builder application will construct a EUMETSAT SIP, from the MTG SIP and NetCDF files for archival.

In the future, the possibility of converting all legacy products into the EUMETSAT SIP format will be considered. This would be an attractive option since it would streamline the software that is required for the archival of products, effectively reducing the number of frontends to one. At the same time, this would guarantee that all EUMETSAT products are archived in a consistent format, this would be beneficial to users of EUMETSAT data because it will support interoperability between EUMETSAT and other organisations.

REFERENCES

- [1] Reference Model for an Open Archival Information System (OAIS) (CCSDS 650.0-M-2) <https://public.ccsds.org/pubs/650x0m2.pdf>
- [2] Earth Observation Metadata profile of Observations & Measurements (OGC 10-157r4, aka EOP O&M), Version 1.1 <http://docs.opengeospatial.org/is/10-157r4/10-157r4.html>
- [3] Space data and information transfer systems — XML formatted data unit (XFDU) structure and construction rules, ISO 13527:2010 <https://www.iso.org/obp/ui/#iso:std:iso:13527:ed-1:v1:en>
- [4] Standard Archive Format for Europe Control Book Volume 1 http://earth.esa.int/SAFE/download/Specifications/PGSI-GSEG-EOPG-FS-05-0001-02-01_CoreSpec_v2.4.zip

FROM HPC TO HYBRID CLOUD COMPUTING

Sébastien Dorgan, Vincent Gaudissart, Stephan Aimé

CS SI France

ABSTRACT

Current and forthcoming Earth Observation (EO) missions are steadily increasing volume, delivery rate, degree of variety and complexity and interconnection of data. This poses challenges addressing the entire EO data lifecycle, from data collection, storage, management, dissemination and access to their processing, analysis, exploitation, integration and delivery to final consumers. This leads to unprecedented opportunities to support and empower new types of user applications, and to develop a new generation of user services[1].

To support the actors of the Earth observation world: students, researchers, institutions, little or big companies, CS SI has designed a cloud computing and data management platform covering all services level: **IaaS** (Infrastructure as a Service), **PaaS** (Platform as a Service), **EO WS** (Earth Observation Web Services), **SaaS** (Software as a Service), **DaaS** (Data as a Service).

Index Terms— Big Data, HPC, Cloud, Container, Software-defined Networks

1. INTRODUCTION

Challenges stemming from such an increase of volume, velocity and variety of data drive the urgent need of new processing concepts, which shall ensure the necessary power but also scalability and elasticity to actually exploit those data.

Furthermore, new requirements from EO data user communities are emerging, and in particular to be able to easily integrate their own processing, manipulation and analysis tools into harmonised frameworks (platforms), which on their side should provide basic processing functionalities, like e.g. efficient data access, parallelisation methods for massive processing, load balancing, etc. Platform users aim at integrating their own processing tools in a seamless and easy way, possibly only requiring a simple configuration to make it become part of a common processing framework, and avoiding software changes and/or development of additional interfaces/components for the sole purpose of their integration and deployment.

To face this challenges we can be helped by the raise of the cloud computing, the development of many new computing patterns under the banner of the Big Data technologies and a variety of libraries and toolboxes for Remote Sensing image processing currently available to EO users to support and facilitate their processing needs.

However we do not forget that most of these libraries and toolboxes has been designed with traditionnal computing in mind and can be arduous to integrate into a Cloud Big Data environment. Additionally a lot of organisations have also invest a lot of money to create processing chains targetting traditionnal High Performance Computing Centers.

A conclusion is to say there is no unique Big Data or HPC Framework to address all computing patterns (Map/Reduce, Streaming, Directed Acyclic Graph ...) and all data types: Satellite Imagery, IOT data, Social network stream ...)

That is why a modern EO computing platforms should be able to combine efficiently Big Data and legacy computing patterns on hybrid on premise and cloud computing infrastructure.

The purpose of this document is to describe the solutions proposed by CS SI to build a such processing platform.

2. OVERVIEW

The design depicted in figure 1 of CS SI cloud computing platform is organized in 4 layers:

- > **IaaS** layer designed to provision computing infrastructures for many public and private cloud computing platforms
- > **PaaS** layer to provide highly available and scalable hybrid HPC and computing platforms on demand
- > **EO WS** layer to facilitate creation, deployment and orchestration of **OGC WPS** services
- > **SaaS** layer to centralise data storage, discovery, visualization, multi dimentionnal analysis and dissemination

This 4 computing and analysis layers are completed by an orthogonal INSPIRE compatible **DaaS** layer compliant with all geospatial standard data management API: **OGC CSW**, **OGC WCS**, **OGC WCPS**, **OGC WMS**, **OGC WMTS**, **OGC WFS**, **OpenSearch**.

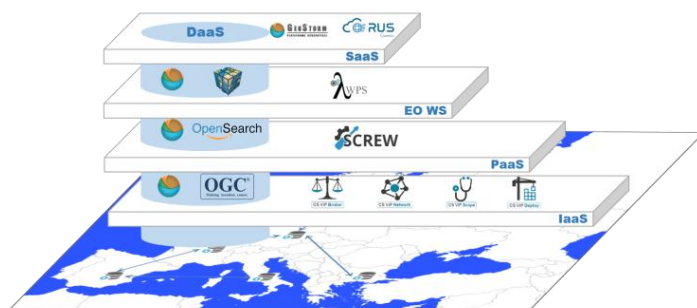


Fig. 1. Design overview

3. IAAS LAYER

A multi-cloud strategy allows to make sure to always have the right offer, to benefit from a maximum of flexibility and to not create dependency on a cloud vendor [2].

For this purpose CS SI has developed **CS ViP** (Critical System Virtual Platform) a multi IaaS system for interfacing with many of the popular cloud service providers using a unified API. CS SI has

been built over cutting-edge devops, monitoring and remote desktop illustrated by figure 2.

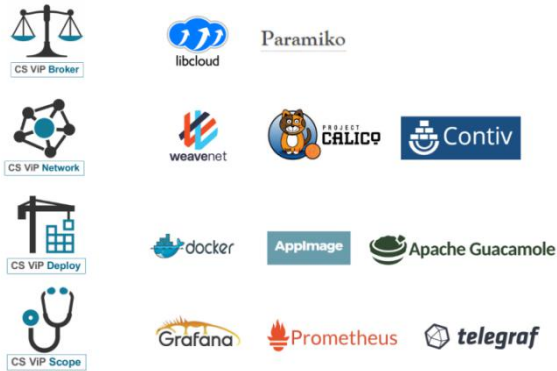


Fig. 2. CS VIP Technology map

CS ViP is divided in 4 components:

- > **CS ViP Broker** to provision virtual machines and storage capacity. With Broker no one care about cloud provider specific terminologies to provision your servers. Ask Broker some resources (cpu, ram, disk) and it will find the virtual server that best fits them among all cloud accounts registered.
- > **CS ViP Network** to create virtual networks interconnecting applications components. Network does not rely on any cloud vendor technology but creates networks overlay completely independently of the underlying cloud network fabric. Beyond that Network can create Software Define Networks (SDN) over many cloud providers that means that distributed applications can be deployed on different cloud providers they will see them each other as if there were on the same local network.
- > **CS ViP Deploy** to package and deploy applications on any cloud provider independently of the hypervisor and the operating systems available. Deploy comes also with a full web desktop solution to manage graphically virtual servers without any additional plugin and give the feeling to work locally.
- > **CS ViP Scope** to monitor resources and applications. Scope collects all metrics in a time series database, provides a query and analysis interfaces and nice charts to monitor system health. Scope can be easily extended via a plugin mechanism to add applications specifics metrics.

CS ViP is used operationally to manage ESA RUS project user environments.

4. PAAS

To face the 5 V challenge (Volume, Velocity, Variety, Variability and Value) on affordable commodity clusters many Big Data frameworks appeared in recent years. Without special cautions the choice of a Big Data framework is necessarily divisive with regards to the others. Moreover a large valuable code database exists for traditional HPC which will be not anymore available.

To reconcile these ecosystems CS SI has designed **SCREW** a PaaS system providing on demand computing platforms combining major Big Data Frameworks: Spark, Hadoop, Ignite... with traditional HPC framework MPI and batch scheduling using DRMAA (Distributed Resource Management Application API) standard.

Technology involved in SCREW are depicted by the figure 3



Fig. 3. SCREW Technology map

Resource management is the central part of the system this feature is bring by **MESOS** a datacenter resource manager capable of dynamically and fairly distributing computing resources to a wide range of computing frameworks using the 2 level scheduling method with the Distributed Resource Fairness (DRF) algorithms[3]. MESOS is capable to manage resources of traditional computing servers but also modern GPU like co-processors necessary to train and run Deep learning systems.

To deal with HPC applications SCREW rely on MPICH an MPI (Message Passing Interface) implementation compatible with MESOS and a MESOS connector DRMAA implemented by CS SI in the frame of a R&T about resource scheduler hybridization conducted by the CNES.

To deal with Big Data computing 3 complementary Big Data frameworks have been selected: Apache Hadoop, Apache Spark, Apache Ignite.

Hadoop was the first open source implementation of the Map/Reduce paradigm and it is the most widely adopted Big Data frameworks. It opens to a vast ecosystem of tools to deal with various computing paradigms: Apache Pig to manage dataflows, Apache Storm for real time analysis, Apache Flink to manage large Directed Acyclic Graphs (DAGs) of operations ...

Spark can be seen as an in memory accelerated version of Hadoop Map/Reduce and it is also particularly efficient to create streaming applications. As Hadoop, Spark is very mature and widely adopted.

Ignite is a versatile framework managing many distributed computing paradigms: Map/Reduce, Even, Distributed closure, Data stream It can be used as an Hadoop accelerator and it can complement Spark to facilitate job communication and synchronisation over the computing grid. Ignite provides also a programming language agnostic distributed in memory key/value store powered by an SQL distributed engine offering an ACID (Atomicity, Consistency, Isolation et Durability) transaction mechanism. Apache Ignite can therefore be considered as the first data management component proposed by SCREW.

Data management has been also completely upset by the raise of the Big Data, the relational paradigm leader for more than 30 years have been challenges by schemalless and scalable approaches under the banner of NoSQL.

HDFS (Hadoop Distributed File System) is the companion persistant key/value data store of Hadoop. HDFS is widely adopted, robust, scalable, fault tolerant and perfectly fit with Map/Reduce computing systems.

In complement of Ignite and Hadoop, SCREW proposed **Apache Cassandra** another widely adopted, robust, scalable and fault tolerant NoSQL system but implementing the Column store paradigm. Compared to key/value stores where the strength is the

simplicity Column stores are better fitted to manage complex data structure and offer more powerful and faster query languages. In a column store database data are organized in columns instead of row in a relational model that bring them more tolerance to the modification of the data structure.

Beyond that we do not forget that numerous applications rely on relational database cannot be excluded from the scope of SCREW and that news application could really benefit of the holistic concepts and strong consistency of a relational model. That's why SCREW proposed a NewSQL database **Postgres-XL**. NewSQL DBMS (Data Base Management System) have been developed to bring to relational database the scalability of NoSQL database. Postgres-XL is fully compatible with PostgreSQL and all its extensions including **PostGIS** the most widely adopted Geographical Information System database.

Last but not least SCREW is not only capable to schedule transparently computing frameworks but it is also capable to orchestrate micro services[4]. A micro service orchestrator let user focused on business problems and manages orthogonal concepts such as service discovery, scalability, load balancing, health monitoring, metrics. The micro service orchestration system of SCREW rely on Marathon a production-grade container orchestration platform for MESOS. This micro service orchestration feature is a strong point because it allows SCREW to deploy the computing intensive components and all the peripheral services barely transparently on the same infrastructure.

A simplified version of the SCREW platform embedding only Ignite and MPICH is deployed for the **RUS project** developed by CS SI and conducted by **ESA** for the **European Commission**. The purpose of RUS is to offer private distributed development and validation environments on demand to students, scientists, and little companies interested by evaluating Copernicus data.

The complete version of SCREW is under development for a first application, with the aim to gradually manage resource allocation from a private commodity cluster to the **CNES HPC** cluster and to public Cloud environments..

5. EO WS LAYER

Lambda WPS is an **OGC WPS** serverless framework that has been designed in the frame of the GSTP Big Data Raf contract with ESA and that will be deployed on future **DIAS** platforms. Lambda WPS is build around docker and Marathon for the resource isolation and the service orchestration and on the SCREW platform and the DaaS layer for processing and management.

A serverless framework allows focusing on business logic instead of dealing with orthogonal concerns such as scalability and high availability. Concretely a user of Lambda WPS submits his source code to the system. This code is converted in OGC WPS micro-service and launched on a SCREW platform by the Lambda WPS service. The health and the load of the service are managed by the Lambda WPS service that restarts the OGC WPS service if it is unhealthy, instantiate another service and managed the load balancing if the service is overloaded. The relationships between Lambda WPS and the SCREW platform are illustrated by figure 4.

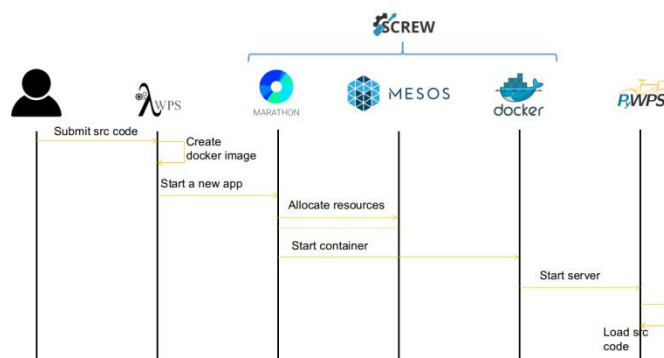


Fig. 4. Lambda WPS workflow

6. SAAS AND DAAS LAYER

The proposed platform will be based on a CS SI, existing product so called **GEOSTORM**.

GeoStorm (stands for GEO Services plATfORM) is a geospatial platform offering storage, discovery, visualisation, multi dimensionnal analysis, processing and dissemination capabilities for many kinds of geo-information. It has the ability to ingest various sorts of geophysical data (Geographic, Hydrographic, Meteorological and Oceanographic) and merge them into layers that are portraying in a seamless mode. It also brings the capacity to host business processes in whatever thematic an run them on an HPC cluster or on a SCREW platform using DRMAA connector. GeoStorm provides the SaaS layer with a cloud ready multi-tenant full web application designed to facilitate the adoption of geospatial information platform and centralise expertise and knowledge in a central product.

It also provides the DaaS layer providing an **INSPIRE OGC CSW** and **OpenSearch** compliant catalog, various dissemination services compliant to OGC WMS, WMTS, WCS, WFS standards and several data storage interfaces: WebDAV, FTP, Swift.

GeoStorm is operationnally deployed on many projects led by CS, in contexts as diverse as civil security (central data aggregation platform, flood management system for various cities...), business intelligence (monitoring and geolocation of news feeds for the French High Committee for the Civil Defence, statistics ...) or spatial world (ongoing integration to the ESA EOCloud platform, s2Agri prototype, integrated into different project for Telespazio, the EU Satellite Center...).

In addition to these existing features CS SI is developing new interfaces to distribute co-registered data as multidimensional arrays known as a **Datacubes**.

Until now EO image data was stored in a file system or in an object storage, could be selected geographically, temporally and filtered by metadata using an INSPIRE compliant catalog and downloaded in their original packaging. To deal with a large number of data sources this approach is not optimal because according to their origin the packaging and the geometry of the products differ and it is up to the user to know how to extract and use data instead of focusing on added value analysis.

This way of distributing data have been a brake for a wide adoption of multi-source and time series data analysis. **Datacubes** are a simple and appropriate alternative to remove these constraints.

The concepts of Datacubes is to stack image tiles on the same geometry, making them then easy to query to retrieve time series of data on a specific geographic location formatted in familiar and ready to compute multidimensional arrays.

Nowadays a large offer of array databases implementing Datacubes exists: Rasdaman, Scidb... a complete review of open source and proprietary implementations is available on the Research Data Alliance web site

To integrate a Datacube into GeoStorm data access layer we have chosen to deploy **Rasdaman** an open source data cube implementation. Rasdaman has been chosen because:

- > it is open source with a GPL v3 license compatible with GeoStorm license
- > It is one of the fastest array database
- > It is official **OGC WCS/WCPS** Reference Implementation
- > It is mature and battle proven: Rasdaman has existed and evolved for more than 20 years and has been successfully deployed on very large environments.

However, finding a good array database technology is not the main challenge when creating a multi-source Datacube. Indeed, these databases presuppose that the ingested data is spatially coherent, i.e. the tiles are projected on the same grid and images are stackable. Unfortunately it is not the case when dealing with multi-source data coming from different satellites and/or different sensors and data has to be co-registered before being ingested in the Datacube.

An exemple of **co-registration** of data coming from different type of sensors is described by the methodology developed by the Munich Technical University in collaboration with the **DLR**[4]

7. CONCLUSION

As shown in this paper, **CS SI solution** is based on state of the art and innovative technologies and makes use of open source software for avoiding any vendor lock-in. Thanks to full compliance to widely adopted open standards, CS SI cloud stack is interoperable to any cloud environment and by the way with upcoming **Copernicus DIAS** platforms and is also a core solution for ensuring their performance.

As a precursor, the **Research and User Support Service (RUS)** that has been running since July 2017, demonstrates the benefits of **CSSI cloud solution** for Earth Observation data processing. **RUS** aims to facilitate the uptake of Sentinel data providing a free access to Copernicus data, cloud processing facilities tailored to user needs, training sessions and materials. The RUS Service is financed by the **European Commission** managed by **ESA** and operated by **CS SI**.

8. REFERENCES

[1] Laurent Probst, Laurent Frideres, Benoît Cambier, PwC Luxembourg & Jean-Philippe Duval, Morgane Roth, Camille Luda, PwC France “Big Data in Earth Observation”, Business Innovation Observatory, European Union, February 2016.

[2] Abhishek Verma, Luis Pedrosaz Madhukar Korupolu, David Oppenheimer, Eric Tune, John Wilkes, “Large-scale cluster management at Google with Borg”, Google Inc, <https://research.google.com/pubs/pub43438.html>, 2015

[3] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, Ion Stoica, “Dominant Resource Fairness: Fair Allocation of Multiple Resource Types”, University of California, Berkeley, https://www.researchgate.net/publication/228950060_Dominant_resource_fairness_Fair_allocation_of_multiple_resource_types, January 2011

[4] Derrick Harris, “Introducing open source DC/OS: The best way to run containers”, <https://mesosphere.com/blog/open-source-dcos/>, April 2016

[5] Sebastian Türmer, “Automatic Registration of High Resolution SAR and Optical Satellite Imagery in Urban Areas”, http://elib.dlr.de/60409/1/DA_Tuermer.pdf, April 2009

DOCKER USED ON SPATIAL GROUND SEGMENT

Christophe Baroux¹, Jean-Christophe Dislaire², Yanis Lisima²

¹ Docker, ²Thales

ABSTRACT

The new Spatial Image Ground Segments need the use of Big Data technologies instantiated on different types of platforms. Consequently, their development and exploitation requires specific features:

- Portability: application must run near where the data are stored. The new deployment tools have to significantly decrease the impact of a new version in term of unavailability (Reduction of the Time To Market) and the use of different kind of platforms (Clouds, legacy Hardware ...),
- Scalability : application must adapt to use huge volume of data,
- Performance and High Availability : use of micro virtualization technology and micro services architecture,
- Networking: adaption to highly distributed data application.

Processing are carried by technologies like Big Data (ex : Hadoop), other characteristics mentioned above are using technologies provided by Docker.

Index Terms - Portability, Scalability, High Availability (HA), DevOps, micro-virtualization, micro-services, Reduction of Time To Market, Docker, Agility, L2PF, EUMETSAT,

1. INTRODUCTION

The spatial ground segment applications are characterized by huge data volume, network distribution, archiving, cataloguing. Our experience gained during applications development confirmed that Docker is particularly adapted. This paper will detailed the advantage of Docker on following aspects :

Docker is one of the major DevOps Tools, recently developed but already widely used by industries requiring micro-virtualization. The success of this technology has been demonstrated through its intensive use by GAFSA companies.

Due to the specific requirements of new Spatial Ground Segments, Thales developed L2PF (Level 2 Processing Facility) project with this technology. This project aims at

offering to EUMETSAT a Big Data Processing Platform for a new generation of meteorological data.

2. FROM VIRTUAL MACHINES TO CONTAINERS

To solve the point concerning data volumetry, Docker proposes container technology. A container is an isolated environment running over an operating system. All containers use shared resources from the host operating system. A virtual machine is a complete emulation of a computer with all its resources and its own virtualized hardware. A container is an isolated environment running over an operating system. All containers use shared resources from the host operating system.

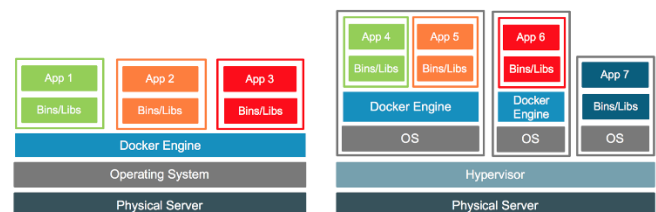


Figure 1: Containers vs Virtual Machines

Consequently a container is much lighter than a VM (hundred of MO vs than some Giga) and offer on intrinsic way a deployment support particularly adapted for application realized for micro-services.

Considering the possibility to install a level of Docker (Docker Engine) on all plate-forms (Eg Linux, Windows), it allows a guaranteed portability. An application working on the developer server will work on same way than on production either the physical infrastructure nor the virtual infrastructure (or mixed).

As showed on previous Figure 1, for better performances and reducing VM licensing cost of the VM, Docker Engine could be installed directly on the Operating System (left part of the Figure 1). For more classical Cloud environment it is possible to install Docker on the VM kipping software applications installed as usual without Docker (right part of the Figure 1).

2.1. Portability

These applications need to be processed on different servers. Thanks to Docker, the project L2PF has been able migrating its development platform from one cloud to another one

with minimal cost and effort. The validation and production platforms are installed directly on dedicated server with just operating system to maximize performance.

As already mentioned, containers are ten times smaller than virtual machines as they only contain libraries required by the embedded application. Docker lowers requirements for running containers to only the host kernel.

2.2. Performances

Due to data volume, but also processing requiring more resources, performances is a key point.

Docker provides better performance than virtualization. Running containers on hardware is like running direct process. Virtualization adds consequent overhead in order to emulate computers.

Docker containers are usually up and running in less than 1s. Using Docker infrastructure allow using more efficiently resources.

2.3. Scalability

Another characteristic joining spatial ground segment applications to docker is the scalability. Some applications are design to run as singleton on systems. In order to build an infrastructure able to serve great loads you need to create several instances of virtual machines and use load balancing and shared storage. Docker simplifies this scalability by allowing several containers of the same application, each running as singleton, on the same system. Docker optimize resources with its storage system by reducing data overhead of containers to the least needed writable data. This is showed on Figure 2 How containers extends scalability of applications, and explained on reference [1].

By using Docker Swarm you extend your application scalability to a cluster of physical or virtual machine and become able to provide more computing capabilities.

To absorb intense traffic usage L2PF is capable of deploying more instances of FTP servers with minimal cost and maximum reactivity.

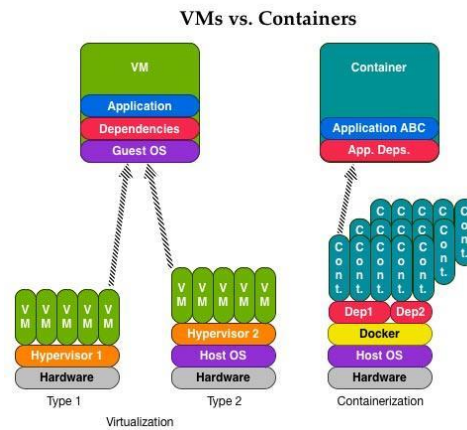


Figure 2: How containers extends scalability of applications

2.4. High Availability

The solution must ensure that system will continue to operate despite serious incidents or disasters. Docker makes HA easier. It provides internal features with simple usage in order to deploy highly available systems. A load balancer allows several containers to deliver the same services. Swarm a native cluster orchestration allows to easily scale number of nodes to up or down.

2.5. Networking

Docker networking gives the same isolation level as virtualization only with many additional features. Docker networks are easier to manipulate. Default networking is internal. Containers expose ports externally only when necessary. L2PF use Pipework [2] to allow containers to use directly the virtual interface from the host in order to guarantee maximum performance.

3. BIG DATA FRAMEWORKS ALREADY DOCKERIZED

Popular Big Data framework and related tools are already running on Docker:

- Kafka for message queuing and buffering,
- Apache Storm for distributed computing, Elasticsearch for indexing,
- Apache Zookeeper for distributed coordination and MooseFS for storage.

This highly scalable distributed software need an infrastructure capable of scaling the same way. Docker is the perfect answer.

L2PF for example deploys about seventy containers on eight physical servers in a Swarm Cluster. Dell FD332 hosts (Storage Hosts) are dedicated to storage type container. Dell FC630 serves processing and other containers. These hosts are mounted in three FX2 chassis. L2PF uses two main networks for the communication between containers: Production and Private. Bandwidth is guaranteed at twenty gigabits per seconds. The MooseFS storage system serves seven terabytes to all containers.

L2PF only use container services : DNS, LDAP, NTP, Jenkins, HA Proxy, Elastic Stack, MooseFS for infrastructure.

Zookeeper, Kafka, X2GO, FTP are running as applications. L2PF has rebuilt more than fifty images in order to maintain packages updated.

Due to requirements from big entities, Docker technology is progressing very quickly. It is why it's so important to dedicate time to follow Roadmap. For example for L2PF, the choice of "dockerise" Basic Services of NTP type was questioned during development. Indeed, Docker was criticized for not being secure (enough), which led to a hardening of the ecosystem and therefore to NOT authorize the containerized software to access low level of resource using Root. It is important to follow the Roadmap but also to participate and follow the ecosystem.

4. Enterprise ready

Docker Inc supports three product lines:

- The Open Source MOBY project [3] is a joint effort between Docker and a community of 2500 developers, to build and deliver an innovative open containers platform.
- Docker Community Edition [4] is a free and open source version of Docker, available for development
- Docker Enterprise Edition [5] is a commercial product designed and built for production, to run enterprise applications built as containers, and deliver a containers platform known as CaaS (Container as a Service)

Docker Enterprise Edition (Docker EE) is designed for enterprise development and IT teams who build, ship, and run business-critical applications in production and at scale.

Docker EE provides an integrated, tested, supported, and certified platform for apps running on enterprise Linux or Windows operating systems and Cloud providers. Docker EE is tightly integrated to the underlying infrastructure to provide a native, easy to install experience and an optimized Docker environment. Docker Certified Infrastructure,

Containers and Plugins are exclusively available for Docker EE with cooperative support from Docker and the Certified Technology Partners.

The Docker Enterprise suite is made of 3 products from Docker:

- The Docker engine, installed on every node of the cluster, managing the deployed containers, and controlling the overall Docker cluster.
- Docker Universal Control Plane, a Docker cluster management layer, in charge of locking down the cluster with added security, access control, scoping and multi tenancy. UCP also bring a web based graphical management interface and integration points with external components like LDAP/Active Directory, PKI, Log management system.
- Docker Trusted Registry, a private version of Docker HUB, that allows the management of Docker images, quality control of images, Vulnerability scanning.

Customers experiences are detailed on Docker site [6].

5. CONCLUSION

Our experience in Docker use shows that this technology is indicated for spatial ground segment applications. Docker will bring outstanding improvement in distributed application project in particular in the area of Portability, Scalability, Performance, and High Availability as well on two others aspects not developed in the paper:

- Interaction with customer during development using rapid prototyping for validation (Agility)
- Improved way of including security during the project and final applications (limitation of micro-containers life expectancy, managing images registry, continuous integration/validation and authentication).

The Docker technology is an open ecosystem for building, shipping and running distributed application at the edge of research already applied in industry.

6. REFERENCES

- [1] Available : <http://sattia.blogspot.fr/2014/05/docker-lightweight-linux-containers-for.html>
- [2] S. Leathers, «Advanced Docker Networking with Pipework,» [En ligne]. Available: <https://opsbot.com/advanced-docker-networking-pipework/>
- [3] Moby Project,» [En ligne]. Available:

<https://mobyproject.org/>

[4] Docker, «The Docker Platform,» [En ligne]. Available: <https://www.docker.com/community-edition>.

[5] Docker, «Docker for the Enterprise,» [En ligne]. Available: <https://www.docker.com/enterprise-edition>.

[6] Docker, «Docker Customers,» [En ligne]. Available: <https://www.docker.com/customers>.

JUPYTEP IDE AS A CONCEPT OF INTEGRATED DEVELOPMENT ENVIRONMENT FOR EO DATA CLOUD-BASED PROCESSING SOLUTIONS

Daniel Zinkiewicz

Wasat Sp. z o.o.
Warsaw, Poland

daniel.zinkiewicz@wasat.pl

Jacek Rapiński

University of Warmia and
Mazury, Olsztyn, Poland

jacek.rapinski@uwm.edu.pl

Michał Bednarczyk

University of Warmia and
Mazury, Olsztyn, Poland

michal.bednarczyk@uwm.edu.pl

ABSTRACT

The idea behind Jupyter IDE is based on building Jupyter notebook Integrated Development Environment for EO data processing. It will be an extension of Jupyter software ecosystem with customization of existing components for the needs of EO scientists and other professional and non-professional users. The approach is based on: a configuration, customization, adaptation and extension of Jupyter, Docker and Spark components on the EO data cloud infrastructure in the most flexible way; an integration with accessible libraries and EO data tools (SNAP, GDAL, etc.); adaptation of existing WPS-oriented EO services provided by TEPs. The user-oriented product will be based on a web-related User Interface in the form of: an extended and modified Jupyter UI (frontend) with a customized layout, the EO data processing extension, and a set of predefined notebooks, widgets and tools. The final IDE will be targeted at remote sensing experts and other users who intend to develop a Jupyter notebook with reuse of embedded tools, common WPS interfaces and existing notebooks.

Index Terms— Jupyter, IPython Notebook, TEP, WPS

1. INTRODUCTION

Interactive Jupyter Notebooks (IPython Notebooks) are state-of-the-art tools in data science. They allow for the analysis and visualization of data in a new paradigm that sits between flexibility and a power of “back-end” processing services, and interactivity coupled with human-friendliness of “front-end” visualization frameworks. The foundational idea is the exposition of Jupyter frameworks on the ESA-supported EO Exploitation Network of Platforms with web services of the basic REPL (read-evaluate-print-loop) adapted for EO data processing shells that are commonly associated with interpreted languages like Python, JavaScript, Scala or R. These technologies usually differ in terms of the underlying tech stack that gives access to various ecosystems and wealth of libraries, each one with its own strengths and a focus on a particular class of problems. All these heterogeneous technologies can be used in an easy way for EO data

processing and product elaboration with use of the extended and adapted Jupyter framework customized for EO data and platforms purposes in the form of Jupyter IDE.

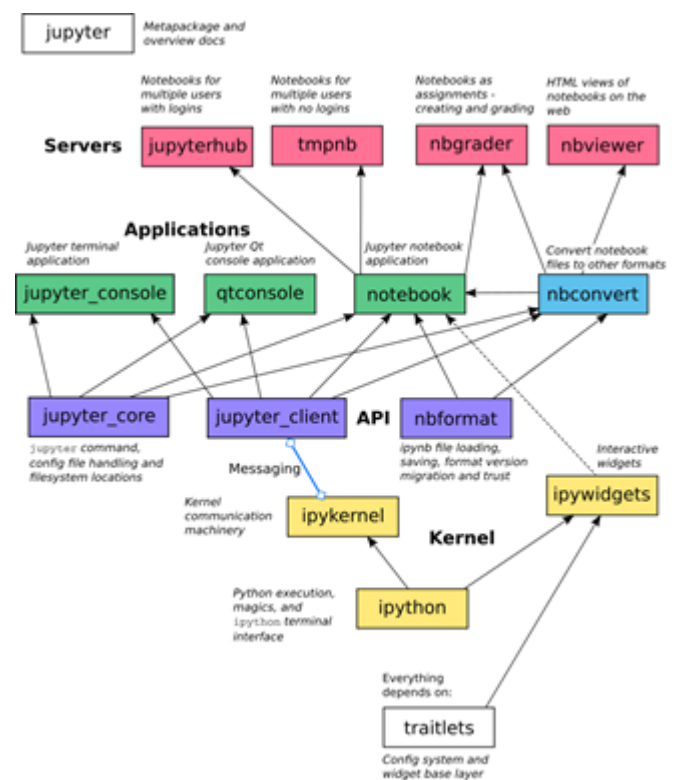


Fig. 1 A high level visual overview of jupyter relationships.

Jupyter is a fairly complex environment (**Fig. 1**). It consists of four basic layers, which are:

1. Kernel - the basic runtime layer responsible for running Python scripts
2. API - a layer that connects the Kernel layer to applications layer, providing communication, file system support, and Jupyter commands

3. Applications - the layer responsible for handling applications via the command line, terminal and notebooks as well as converting notebook files to other formats

4. Servers - a layer that serves notebooks in the HTML format, also providing access to notebooks for many users

The basic functionality is provided by the IPython interpreter, but Jupyter's most useful feature is the ability to create notebooks. They are a convenient tool for creating applications using multiple programming languages and having their own user interface.

Jupyter is also extensible, so one can equip it with his own functions. For this reasons, it is a good basis for JupyTEP IDE development. It is also very important in this case that Jupyter is distributed under the open-source licence, which allows for the modification and use of the project in its entirety.

2. OBJECTIVES

JupyTEP IDE will be designed as a part of a network of exploitation platform software and the EO data cloud infrastructure. In order to cope efficiently with the problems present in EO data cloud processing, the following general objectives for JupyTEP IDE have been formulated:

- EO tools integrations: JupyTEP IDE will allow for significant leveraging of the available EO developer tools (readers, mappers, libraries) to perform EO data specific tasks
- Interoperability: JupyTEP IDE will extend an ability to integrate the notebooks in existing solutions/platforms, providing an added value to the existing environment
- Multi-user/Multi-tenancy: JupyTEP IDE will be served as a web-based solution to multiple users, or - going further - to different groups of users (each as a "tenant") with a proper isolation and mutualisation of cloud resources
- Scalability: JupyTEP IDE will provide an ability to process a growing amount of EO data by an expanding community of developers and scientists
- Parallelization: linked with scalability, allows for making use of a large amount of resources to perform more work in the same time or the same work in a shorter period.

3. CONCEPT

JupyTEP IDE will be mainly realized by the development of new Jupyter extensions or plugins, which enable end users to control the behaviour and the appearance of the Notebook application. The extensions capability can vary from being able to load notebook files from GitHub repositories, Google Drive, or PostgreSQL server, presenting the Jupyter notebooks in the form of a grouped list of available functionalities or services, to just adding a convenient UI element (e.g. button) or a keyboard shortcut for an action frequently performed by a user.

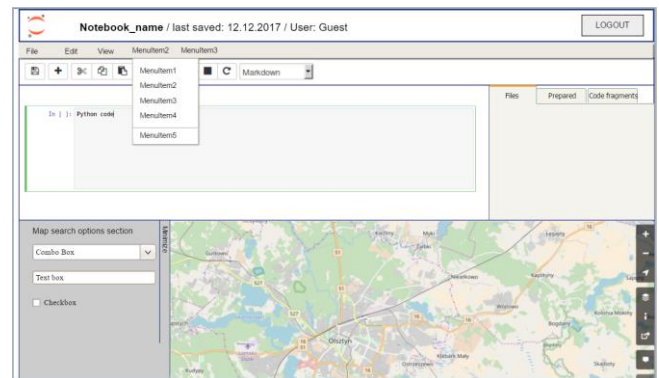


Fig. 2 Mock-up of JupyTEP IDE user interface.

The approach to implement JupyTEP IDE is to provide the minimal sensible default with an easy access to a configuration for extensions to modify behaviour of Jupyter environment. Those extensions will be composed of many pieces, but - from a technical point of view - they will be based on a JavaScript part that lives on the frontend side (e.g. the UI, written in JavaScript), and a part that lives on the server side (written in Python). During the development of IDE part the focus will be on the JavaScript side, customization, and adaptation of a user environment for the needs of EO scientists and developers.

All requirements of the exploitation platforms (mainly TEPs) and interactive notebooks technology will be retrieved to build modern tools for sharing EO tools and algorithms in the form of notebooks outputs and visualisation of EO data processing results and services chains. Well-understood and well-tested standards, technologies, designs and concepts will be taken into account to ensure a fast, safe and cost-effective implementation of notebooks on the EO platforms. The architecture concepts will be based on a solution that allows easy and cost-effective data and services integration in existing platforms infrastructure.

4. MAIN COMPONENTS

In the development phase most of the work will be focused on building IDE for the EO-related development. It will be based on the Jupyter web console and Jupyter-related tools extensions and widgets. Most of the work will be realized as a development of a new EO data processing functionality and an extended Jupyter environment. The entire development work allows to build a Jupyter IDE designated for developers, scientists, professional users of EO data, and to integrate existing EO tools, EO data and the exploitation platforms infrastructure.

Jupyter implemented as a part of JupyterHub environment will be the core of the proposed JupyTEP IDE. It will be extended and modified by adding new EO-related functionalities and sample EO notebooks, and integrated with exploitation platforms outputs, which are available straight from the IDE perspective. Implementation of other potential Jupyter functionalities may be based on components

integrated within JupyterLab, which is an extensible computational environment for Jupyter. With JupyterLab there will be a possibility to create services that meet the workflow needs, but in this stage of JupyterLab development it is very difficult to integrate it with other components. In this case the plan is to include it if JupyterLab will be published on a stable version.

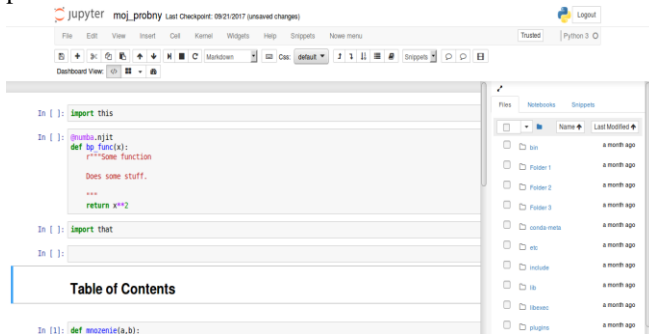


Fig. 3 Integration of Jupyter web UI with jupyter IDE side menu extension

All installed Jupyter IDE components will be extended or redesigned for the EO data and exploitation platform purposes. The Jupyter extension engine is a set of different Jupyter-based components, which contains a collection of extensions that add functionality to the Jupyter notebooks. Another group of reused extensions are layout extensions. The dashboards layout extension, which will be implemented as part of larger Jupyter Dashboards, is a Jupyter Notebook add-on and it allows for arranging notebook outputs (text, plots, widgets, etc.) in grid- or report-like layouts. It saves information about layouts in a notebook document. The extension is a part of the larger Jupyter Dashboards effort which covers: arranging notebook outputs in a grid- or report-like layout; bundling notebooks and associated assets for deployment as dashboards; serving notebook-defined dashboards as standalone web apps. All these functionalities will be adapted for EO data processing and visualisation purposes.

A part of the components will be installed in an early stage of the project and will be available in an alpha version of Jupyter IDE. In the next stage of the project all of them will be extended or redesigned for EO data and exploitation platform purposes. One of them is the Jupyter extension engine, which, as mentioned above contains a collection of extensions that add functionality to the Jupyter notebooks. These extensions are mostly written in JavaScript and will be loaded locally in a browser when needed. Another group of reused extension are layout extensions. The dashboards layout extension which will be implemented is an add-on for Jupyter Notebook and it allows to arrange notebook outputs (text, plots, widgets, etc.) in grid- or report-like layouts. It saves information about layouts in a notebook document. The extension is a part of the larger Jupyter Dashboards effort which covers: arranging notebook outputs in a grid- or report-like layout; bundling notebooks and associated assets for

deployment as dashboards; serving notebook-defined dashboards as standalone web apps. All of these functionalities will be adapted for the EO data purposes.

Another group of components to be integrated with Jupyter IDE and extended for the EO data processing purposes will be different frameworks, widgets and Jupyter-based extensions. The cookiecutter project will be integrated to support getting up to speed with the packaging and distribution of Jupyter interactive widgets.

Beside a customization, it is planned to develop a set of notebooks, which will be available directly from Jupyter UI perspective (Fig. 3). In general, it will be a set of Jupyter notebook scripts or algorithms schemas for the EO data processing ready for use/reuse. Most of them will be based on WPS distributed by different exploitation platforms or EO services providers. In terms of developing the Jupyter notebooks, it is planned to integrate ESA toolboxes as the API-based backend software. ESA SNAP toolbox will play the main role of the processing software and will be exposed in a notebook form that allows intermediate users to invoke the toolbox services through both a native Java-based API and a Python binding (Snappy) format. The developed set of notebooks allows to consume/provide OGC services to intermediate users (service providers or platform users), such as WMS, WCS, WPS, WFS, OpenSearch. In this architecture, the notebook can be exposed as a WxS service or it can be possible to call such a service within a notebook. In addition, a part of the development phase will be allocated to development of visualization tools. Advancing the visualization capability will be achieved with respect to the basic plotting functionality provided by existing libraries (e.g. Python matplotlib, Pivottablejs, etc.) and by making results available on an interactive, customizable widgets or maps (e.g. as per Google Earth Engine).

5. TOOLS AND LIBRARIES FOR INTEGRATION

Existing state-of-the-art open source concepts, technologies, libraries and toolkits will be used to design and develop Jupyter IDE and to assure capabilities for the high data volume information processing, extraction, analytics and fusion. At the integration stage the Jupyter IDE will be ready to adapt technologies used by different platforms and the EO cloud infrastructure by utilization of generic solutions from the big data landscape e.g. HDFS file system; databases like HBase, Cassandra, MongoDB; SQL-like (e.g. Hive, Impala, SparkSQL); machine learning as Spark ML and, quite obviously, the Hadoop framework (with modules like HDFS - distributed file-system; YARN - the resource-management platform responsible for managing computing resources; MapReduce - the programming model for large scale data processing; GeoJinny - spatial Hadoop). Geospatial open source solutions specific for particular platforms will

also be used to integrate JupyTEP IDE with the platform software infrastructure and platform requirements.

```
print("Plotting")
imageData.shape = h, w
resultData.shape = h, w
fig, (ax1, ax2) = plt.subplots(1,2,figsize=(w/1000,h/1000))

ax1.imshow(imageData,cmap="gray")
ax2.imshow(resultData,cmap="gray")

ax1.set_title('Original image')
ax2.set_title('Image of the result')

plt.show()

# del imageData
# del resultData
# %reset out
```

Reading image
Data processing
Plotting

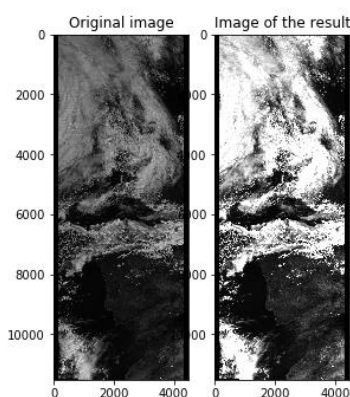


Fig. 4 Processing EO data based on JupyTEP IDE capabilities and integrated EO related tools.

JupyTEP IDE notebooks processing capabilities will be developed, integrated and tested using the OGC open standards like WPS (Web Processing Service) and WCPS (Web Coverage Processing Service) and the best open source implementations (rasdaman, ZOO project, SNAP (**FIG. 4**), GDAL, Orfeo Toolbox, PostGIS, GRASS GIS, GeoTrellis, GeoMensa, GeoServer, etc.). Validation procedures of the developed components will be tested using industry standards and the available EO cloud infrastructure, and operational services hosted by TEPs. Validation and qualification of the developed concept and implemented modules will be tested with the help of the user community utilizing a collaborative and transparent environment.

6. OPEN SOURCE ROLE

The full adaptation and reuse of open-source components will be the main paradigm in JupyTEP IDE development. JupyTEP IDE and Jupyter notebooks as an implemented set of reusable open source components aim for integration in the Network of Platforms Architecture. This should be time- and cost-effective, as flexible as possible, and realized in line with ESA's up-to-date plans. The alternative solutions will be

recognized in view of any technical or other constraints for integration of open source components.

7. SUMMARY

JupyTEP IDE, an example of evolution of the EO tools, integrates most of the aspects of data reception, processing, visualization, archiving, access facilitation, information and knowledge dissemination and management. The IDE allows for achieving these objectives by reducing costs of services implementation and by increasing the developer and science community involvement and responsibility.

8. ACKNOWLEDGMENT

This paper is based upon the ESA-funded project newly launched in the Polish Industry Incentive Scheme:

JupyTEP IDE – Jupyter-based IDE as an interactive and collaborative environment for development of notebook style EO algorithms on Network of Exploitation Platforms infrastructure.

9. REFERENCES

- [1] Project Jupyter, Jupyter Notebook Documentation Release 5.0.0., April 05 2017, Accessed July 2017, <https://media.readthedocs.org/pdf/jupyter-notebook/stable/jupyter-notebook.pdf>
- [2] Project Jupyter, Jupyter Documentation Release 4.1.1.alpha, Accessed November 2017, <https://media.readthedocs.org/pdf/jupyter/latest/jupyter.pdf>
- [3] About the GNU Project. FSF (Free Software Foundation), (2007). <http://www.gnu.org/gnu/>
- [4] Spark cluster on OpenStack with multi-user Jupyter Notebook, September 21, 2015, <https://arnesund.com/2015/09/21/spark-cluster-on-openstack-with-multi-user-jupyter-notebook/>
- [5] Deploying JupyterHub for Education, Jessica Hamrick, March 24, 2015, <https://developer.rackspace.com/blog/deploying-jupyterhub-for-education/>

EARTH OBSERVATION DATA EXPLOITATION IN THE ERA OF BIG DATA: ESA'S RESEARCH AND SERVICE SUPPORT ENVIRONMENT

R. Cuccu^{1,3}, G. Sabatino^{1,3}, J.M. Delgado^{1,3}, J. Van Bemmelen² and G. Rivolta^{1,3}

¹ESA Research and Service Support, ²European Space Agency, ³Progressive Systems Srl

ABSTRACT

The ESA Research and Service Support (RSS) service offers support to ease Earth Observation (EO) data exploitation. RSS users are Principal Investigators, public institutions, SMEs and companies requiring support to progress in their research and development activities. For EO data users it is becoming more and more important to be able to exploit the large time series of ESA's heritage missions data in combination with the data from the Sentinel missions. RSS makes available ad-hoc services which can be easily tailored on the user needs to enable large time series production. Such services comprise: access to a comprehensive and dynamic EO data catalog, provisioning of hardware and software resources for algorithm development, and access to a processing environment for global scale intensive processing.

Index Terms — *EO Big Data access, EO Big Data processing, support, development, EO training.*

1. INTRODUCTION

Support to research, innovation, exploitation and valorization of large multi-temporal datasets requires the set-up of dedicated services and the availability of suitable resources. The ESA RSS [1] is an operational service that, from its first set-up in 2006, has been evolving and enlarging its scope to support the entire EO user community interested in EO data exploitation.

The services provided by RSS can be classified in:

1. EO Data access and provisioning
2. EO Data processing
3. Algorithm/Service development support
4. Training/Mentoring in EO data exploitation and related tools
5. Communication and information sharing.

This paper describes the services offered by RSS, focusing in particular on multi-temporal and multi-mission data support.

2. THE RSS SERVICE

The ESA RSS supports different phases of the research process. RSS makes available e-collaboration environments

to discover and share information (Wiki pages, Forums, Blogs can be created on request), reference and sample datasets, access to a large EO data archive (both ESA and Third Party Missions data), customized Cloud Toolboxes to develop or fine-tune algorithms on selected datasets, on-demand processing environments where new algorithms can be integrated and made available as EO applications for multi-temporal and multi-mission massive processing, and results visualization tools.

2.1. RSS Users

RSS users are scientists, researchers, service providers and SMEs interested in the development of innovative retrieval methodologies or services based on EO data.

Since 2014, RSS has been enlarging its support including also university groups and researchers not necessarily experts in the EO domain, who need guidance and training for EO data exploitation, helping them in understanding EO data applications. Scientific users are supported for the entire duration of their project, from the hypothesis formulation until the results publication. SMEs and companies are supported during the development phase of their innovative project.

2.2. Processing environments

Two types of processing environments are made available depending on the user needs. For multi-temporal and multi-mission massive data processing, the most suitable environment is based on a parallelized and high-performance distributed system, while for early development/testing of algorithms and for processing over limited areas of interest the proper environment is the RSS Cloud Toolbox.

The RSS generic processing platform makes available, through a user-friendly web interface, processing services which are based on the algorithms developed and provided by the Principal Investigators. RSS takes care of integrating the algorithms in the processing platform to create the services. The processing environment allows registered users to configure and run processing tasks, to monitor the status of the task and also the status of the different blocks each processing task is composed of.

Figure 1 shows an example of query result on one of the available altimetry services [2]: the user configures an area of interest in the form of a bounding box, defines start and stop times, and choose a dataset. The query returns all the available products, and the footprints are shown on the map.

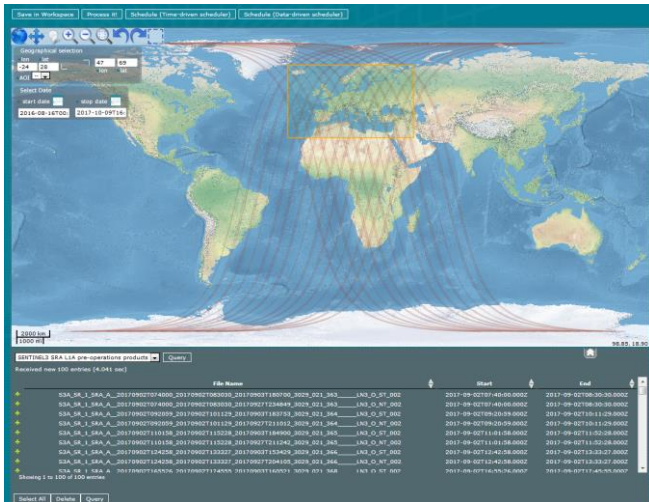


Figure 1. Query result on one of the available datasets of the RSS processing platform. The query returns all the available products intersecting the area of interest in the defined time window.

The user configures the processing parameters and submits the task which will run on the distributed system. The processing flow is shown as a chain of modules which are executed separately on the Worker Nodes (Fig. 2).

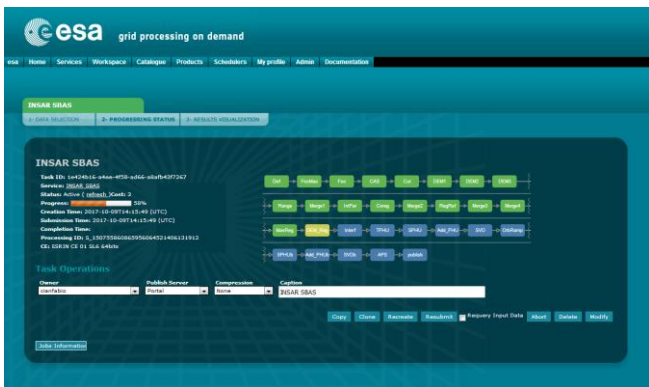


Figure 2. Processing task execution chain.

When the processing task is completed, the results can be optionally downloaded via http from the portal itself or delivered to the users premises, for example via ftp.

The RSS processing platform allows on one hand on-demand processing, which is usually carried out by the user in autonomy, and, on the other hand, scheduled systematic processing, which requires interaction with RSS Team. The current RSS processing platform relies on an infrastructure comprising over 90 processing nodes with a total of about 2.4 TB RAM and 600 CPUs. Such RSS base capacity is on average sufficient to satisfy the processing

requests but it can be easily expanded: when the processing requests increase it is possible to dynamically scale up the resources by seamlessly configuring additional processing clusters located on a generic Cloud Provider Infrastructure or on other HPC systems. Extra processing resources are usually kept only for the time needed to address the extra requests, thus limiting the infrastructure costs.

The RSS Cloud Toolbox [3] is the basic tool offered by RSS to EO researchers and private companies to support the algorithm/service development phase. This tool is a virtual machine based on Linux operating system (both CentOS and Ubuntu distributions are supported) with pre-installed software, powerful enough to speed-up the algorithm development phase. The Cloud Toolbox is completely customisable (additional software, packages and libraries can be installed on it), it is flexible (hardware resources like RAM, CPU cores and Hard Disk can be changed when needed), and accessible from everywhere.

This service is accessible at the following URL <http://eogrid.esrin.esa.int/cloudtoolbox>.

A non-exhaustive list of pre-installed software includes: (i) Sentinel Application Platform (SNAP) [4]; (ii) BEAM; (iii) Sen2cor plugin; (iv) Quantum GIS (QGIS); (v) R-studio. Additional software can be installed on request. The RSS Cloud Toolbox displays its desktop onto users' PCs giving the feeling that the Virtual Machine is locally installed (Fig. 3).

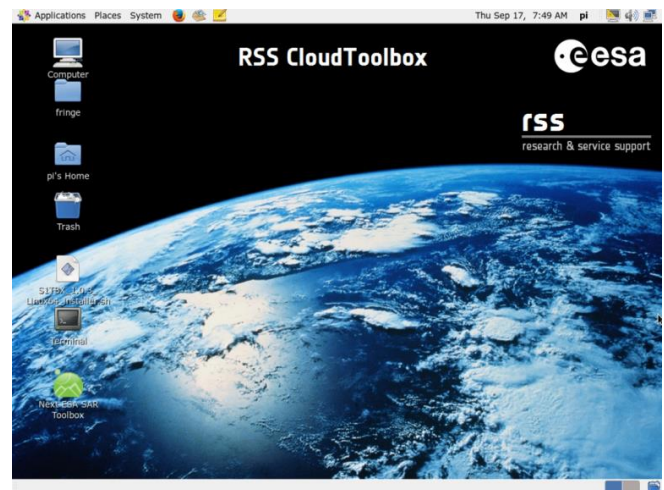


Figure 3. RSS Cloud Toolbox. The graphical desktop is displayed on the user's PC or Laptop through a client-server application (X2Go or VNC).

Users can build their own work environment benefiting from the scalability of the Cloud resources. The user algorithms developed on the Cloud Toolbox, once reached a sufficient level of maturity and robustness, can be integrated into the RSS processing environment, thus bringing the algorithm close to the data either if the scientist plans to run it on massive datasets (multi-temporal processing) or to make it available to the scientific

community as a web application for processing on-demand. An example of processing service which is opened to a community is the ERS/ENVISAT SBAS processing service for interferometry developed by CNR-IREA and made available to registered users on the RSS processing platform.

2.3. RSS support to Business Incubation Centres

RSS has recently started to support a new type of requests coming from the Business Incubation Centres (ESA BICs), from start-ups but also larger service companies with limited knowledge of EO data exploitation. Such type of support is part of the RSS service scope, and it is aimed at providing tools and information to start-ups to foster the development of innovative ideas.

The support to BIC start-up companies includes the provision of tools, resources, information and reference dataset, but also a first level of consulting especially for start-ups approaching EO data exploitation for the first time. RSS delivers such support to promote EO data exploitation and valorisation, to lower the barriers for EO data users and to facilitate the engagement of new types of non-EO users who could benefit from EO data.

3. RSS SUPPORT IN NUMBERS

3.1. Cloud Toolbox support

In the last years RSS has provided support for workshops, courses, university lectures, and to single research groups through the Cloud Toolbox service. Cloud Toolboxes are provided to the users for limited periods (typically from a few months to more than one year). During this time period the users work generally in autonomy. If users need support (both technical and scientific) they can contact RSS Team and will receive assistance typically within 1-2 days.

RSS manages on average about 80 active Cloud Toolboxes and carry out the operations for the creation of the virtual machines, the configuration of the environment, the installation of the software and provisioning of data. RSS can also provide initial training to the users on how to use EO data tools like Sentinel Application Platform (SNAP) or Quantum GIS (QGIS).

3.2. Big Data processing support

With about 300 TB of data stored in-house and access to external repositories (like CryoSat-2 dissemination server, SMOS dissemination server, Copernicus Data Hub, ESA Heritage Mission dataset) allowing on-the-fly data download, the RSS processing environment offers the possibility to access to a very large multi-mission dataset. Besides this virtually unlimited data catalogue, the RSS processing environment provides a HPC distributed system based on Grid and Cloud Technology where processing

tasks are orchestrated and parallelized. Yearly, more than 20 projects needing processing campaigns are supported by RSS, roughly 50000 processing tasks are created, and 1.2 PB of data are processed.

Two recent success stories of RSS support to Luxemburg Institute of Science and Technology (LIST) and German Aerospace Centre (DLR) on multi-temporal data processing are briefly described.

A. Global Flood Record

The main objective of the Global Flood Record is to generate inundation maps of past flood events based on an archive of Synthetic Aperture Radar data: ENVISAT ASAR Wide Swath Mode (WSM). RSS has integrated a dedicated service divided in two processing steps: (i) pre-processing of the SAR data using SNAP Calibration algorithm;

(ii) flood event identification using the algorithm provided by LIST [5]. Figure 4 shows an example of Flood Image (top), Reference Image (middle) and Flood detection (bottom) resulting from the LIST processing.

Currently the processing over Europe (full mission dataset) has been completed and the results have been delivered to the PI for validation.

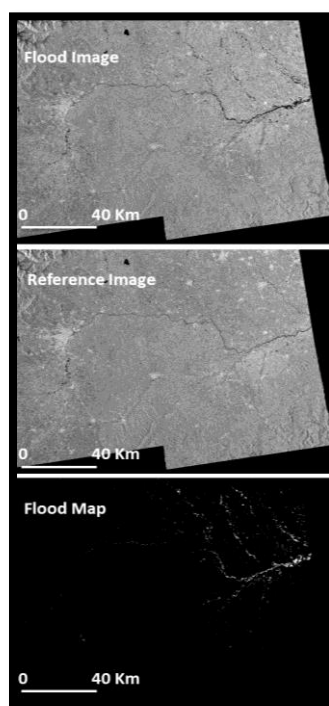


Figure 4. Flood Map Detection – Algorithm provided by LIST.

Based on the output of this validation new inputs will be given by LIST to RSS service in order to extend the processing campaign at a global scale. This work will help to derive inundation risk maps.

B. Urban Extent Map

In this project, the German Aerospace Centre (DLR) has developed a new technique to derive the urban extent maps from 2002 until 2012 for the entire world by exploiting the ENVISAT ASAR WSM dataset. RSS has created a dedicated service to calibrate and geocode the entire ASAR

WSM dataset and to publish the results to the user premises. Runs on different areas of the world are in progress and results are delivered daily. The output is being post-processed at DLR facilities to obtain the temporal-spatial indicators from which urban extent maps for different periods are derived [6]. Figures 5 and 6 show the urban extent maps derived from this study for two African cities: Kampala city (Uganda) and Nairobi (Kenia).

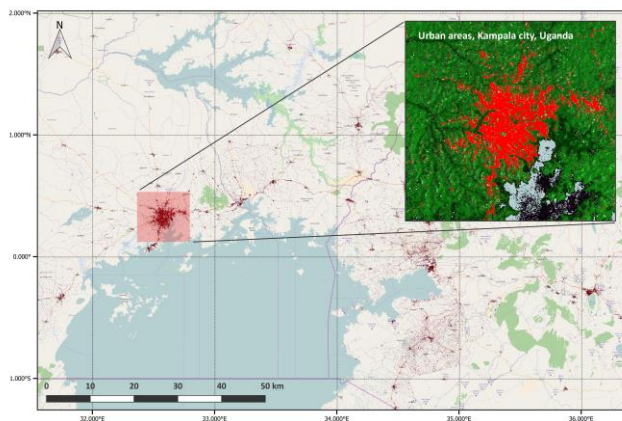


Figure 5. Urban extent map for Kampala city (Uganda) computed from DLR post-processing of ASA WSM L1 images (ENVISAT) processed by ESA-RSS.

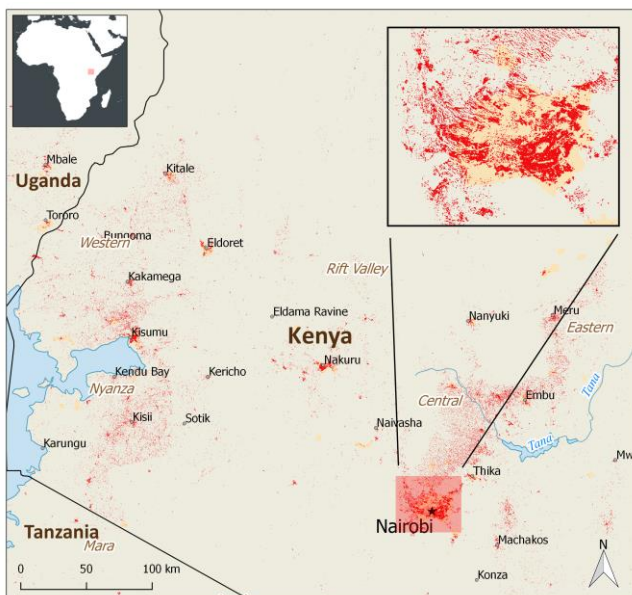


Figure 6. Urban extent map for Nairobi (Kenia) computed from DLR post-processing of ASA WSM L1 images (ENVISAT) processed by ESA-RSS.

4. CONCLUSIONS

RSS makes easier EO data exploitation and valorisation both supporting EO data experts and newcomers. Different services and solutions are made available depending on the users need.

The main objective of the RSS service is to ease EO data exploitation and to sustain research and development productivity by minimizing time and effort that EO data users usually invest in data collection and processing, as well as infrastructure costs for data storage and processing. That allows to free resources on the user side to focus on the development and analysis phase, thus speeding-up production of valid results.

5. REFERENCES

- [1] Marchetti P.G., Rivolta G., D'Elia S., Farres J., Mason G. and Gobron N., "A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support." *IEEE Geoscience and Remote Sensing Society Newsletter*, Vol. 162, pp. 10-18, March 2012.
- [2] J. Benveniste, S. Dinardo and B. Lucas, "SAR Processing on Demand Service for CryoSat-2 and Sentinel-3 at ESA G-POD." *Geophysical Research Abstracts*, Vol. 17, EGU2015-14850, 2015.
- [3] R. Cuccu, G. Sabatino, J.M. Delgado and G. Rivolta. "Enabling SAR data exploitation by processing on-demand", in *Geoscience and Remote Sensing Symposium (IGARSS)*, 2015 IEEE International, vol., no., pp.1476-1479, 26-31 July 2015.
- [4] Y. L. Desnos, M. Fomelis, M. Engdahl, P. P. Mathieu, F. Palazzo, F. Ramoimo and A. Zmuda. "Scientific Exploitation of Sentinel-1 within ESA's SEOM programme element". In *Geoscience and Remote Sensing Symposium (IGARSS)*, 2016 IEEE International (pp.3878-3881). IEEE. July 2016
- [5] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A hierarchical split-based approach (hsba) for automatically mapping changes using sar images of variable size and resolution: ood inundation as a test case", *IEEE Transactions on Geoscience and Remote Sensing* (Submitted), 2016.
- [6] M. Marconcini, A. Metz, T. Esch, J. Zeidler and M. Paganini. "Towards global urban mapping by means of ESA SAR data the SAR4Urban project". *ESA Living Planet Symposium 2016*, Prague, Czech Republic, May 2016.

INTERCONNECTING PLATFORMS VIA WPS: EXPERIENCE FROM THE CTEP/PEPS CONNECTION

S. Clerc⁽¹⁾, N. Gilles⁽¹⁾, M. Paulin⁽²⁾, G. Ceriola⁽³⁾, E. Poupard⁽²⁾, V. Garcia⁽²⁾, Ch. Taillan⁽²⁾,
M. Aspetsberger⁽⁴⁾, S. Barrau-Huguet⁽⁵⁾, Y. Moreau⁽⁵⁾

(1) ACRI-ST, France, (2) CNES, France, (3) Planetek, Italy, (4) Catalysts, Austria, (5) Thales, France

ABSTRACT

In recent years, the concept of platforms offering remote processing services has become the new standard for the exploitation of large data archives.

The WPS standard offers the promise of enabling interconnections between these platforms, allowing users to work seamlessly on several platforms from a single interface.

An experiment with such an interconnection is currently in progress between the Plateforme Pour l'Exploitation des Produits Sentinel (PEPS) of CNES and the Coastal Thematic Exploitation Platform (CTEP, ESA project). This connection will allow CTEP users to execute processing tasks on the Copernicus archive of PEPS.

The authors have in particular developed specific mechanisms beyond the WPS standard to identify the end-user and transfer the processing results.

Index Terms— Exploitation Platforms, Remote Processing, OGC Standards

1. INTRODUCTION

The last few years have seen the emergence of on-line Exploitation Platforms for Earth Observation (EO) data. These platforms offer access to large data archives as well as remote processing services to extract information from the raw data. This concept is progressively replacing the traditional download-based approach which is no longer applicable to the Big Data era [1-5].

The Coastal Thematic Exploitation Platform (CTEP, see [6]) is an exploitation platform devoted to the observation of coastal areas. This service is developed in the frame of an ESA contract and has been operational since July 2017.

The *Plateforme pour l'Exploitation des Produits Sentinel* (PEPS, see [7]) is the French Copernicus Collaborative Ground Segment. This platform offers an access to the full Copernicus data archive (Sentinel 1, 2 and 3 products). Since 2017, the platform is also offering prototype remote processing services for pilot projects in the frame of the “BOOSTER” initiative.

As part of this initiative, an inter-connection between the CTEP and the PEPS is currently under development. In this paper we present the lessons learnt from this interconnection, in particular regarding the use of the Web Processing Service (WPS) standard.

2. THE COASTAL THEMATIC EXPLOITATION PLATFORM SERVICES AND ARCHITECTURE

The CTEP (<https://coastal-tep.eo.esa.int>) offers a variety of services to the community:

- Data catalogue search and discovery service with thematically relevant datasets: Ocean Colour products, coastal satellite altimetry, sea surface temperature, optical observation in coastal areas...
- Remote processing services, including a catalogue of pre-defined thematically relevant processors: atmospheric correction, land/water mask...
- On-line interactive applications such as SNAP, QGIS, Jupyter notebooks, or virtual linux desktops
- Accounting services (user account monitoring and management)
- A processor integration interface to upload, describe and publish new processors (see Figure 1)

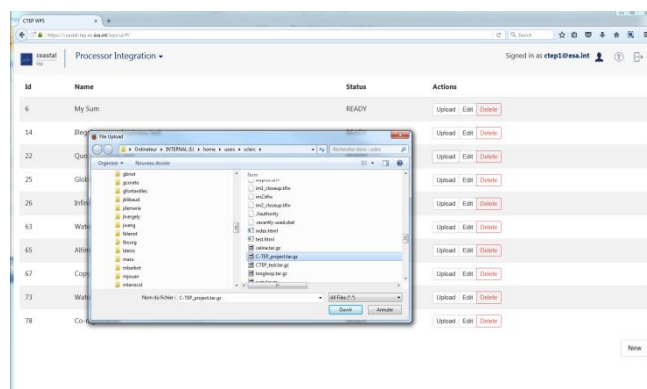


Figure 1: The CTEP processor integration interface enables users to upload software packaged in a compressed archive; the software is then automatically integrated in the platform.

The latter point is a particularly innovative aspect of this platform. It allows users to integrate new processors very easily with a very limited learning curve. Software packaging (based on the creation of a Docker container and WPS metadata) is entirely automatic and accessible from a user-friendly interface which hides the complexity of the integration process. This interface allows users with limited training to integrate a new processor typically in half a day (including debugging effort) without any external support.

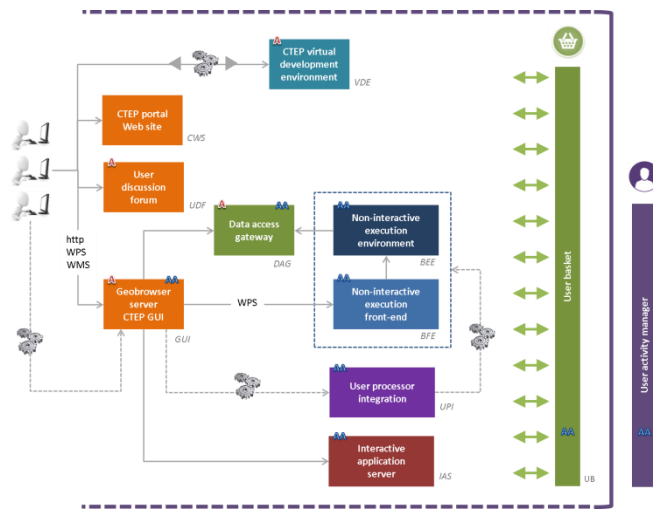


Figure 2: CTEP architecture overview.

The CTEP architecture (see Figure 2) is based on the following main components:

- The Geobrowser, the main User Interface of the platform. It interfaces with the other CTEP components, supports some Web GIS (Geographic Information Service) functionalities, and an interface to manage user data.
- Data Access Gateway which offers catalogue search and discovery services as well as product access management (including support for authorization for restricted access data)
- An original python-based implementation of a WPS server named WISPY. This server defines abstract WPS handler classes, of which the CTEP-scheduler provides a specific implementation. The CTEP-scheduler relies on a dynamic database of processing services. It also offers advanced functions to support user authorization (using Shibboleth) and accounting. The solution implements the WPS 1.0 standard and supports also the “job dismiss” request (part of WPS 2.0).
- The execution environment handles processing requests. It offers functions to manage the computational cluster and to manage input and output data. Input data is copied locally on the computing

- node, while output data and execution logs are copied after completion of the job to the user data storage area.
- The on-line interactive application server based on noVNC technology
- The User Accounting Manager, which manages user resources (access rights to services, CPU time resources...). The component offers a REST API to verify, allocate and consume resources, as well as a graphical interface for user and administrator.
- The processor integration interface collects information to generate a new processor (processor title and abstract, input parameters, execution entry point), and handles the processor upload and integration tasks (including generation of the Docker container and registration in the WPS process database).

All components are developed as Free and Open Source Software (FOSS).

3. PEPS DESCRIPTION AND ARCHITCTURE

PEPS - Plateforme d’Exploitation des Produits Sentinelles - has been developed by CNES to facilitate the usage of Sentinel data by the user communities and thus boost the application development. Close to its users, listening to them to take into account their needs, PEPS offers various access to data and processing capabilities. Starting form 2015, PEPS, as the French “Sentinel collaborative ground segment”, provides through its web portal – <http://peps.cnes.fr> – a full access to all the Sentinel products according to the Copernicus open and free data policy.

With a storage capacity of 14 Pb of data, PEPS provides the following services:

- On-line product user access
- Discovery, view & data loading module
- Basic tools
- Product catalog
- On-demand & automatic processing

The technical solution addresses modularity and scalability, performance, component reuse, communication by services, open source solutions. PEPS’s Infrastructure is deployed on 3 servers completed by an HPSS High Performance Storage Systems, and HPC center (High Performance Capacity).

The ingestion server is connected to ESA’s DHUS (the Data Hub Server for dissemination of the ESA Copernicus Sentinel data access). The server ingests the products on a HPSS (High Performance Storage Systems), and communicates to the distribution server for catalog update. Software component design for the ingestion part is based on an Actor Model.

A web portal provides access to data by search or by navigating on maps. It is developed with Angular JS technology and is compliant with all web browsers.

The product catalogue and search engine are implemented with RESTo (Restful Semantic search Tool for geOspatial). All these services are provided through REST API allowing access from distant application. Based on these same services, a website offers a direct access to users.

RESTo implements different kinds of searches. It complies with the OpenSearch standard with the extensions « GeoSpatial and Temporal » and « Extension for Earth Observation », thus enabling searches with spatial and temporal criteria. Users can also ask for a search expressed in natural language. The query is analysed by the semantic analysis module of RESTo.

PEPS provides to users some ready-to-use geospatial preprocessing services following the WPS standard. These services allow users to remotely launch a processing task through the web interface or to call them directly through an API. The task can be executed either on the processing server or on the CNES's HPC shared server.

The standardized processing mechanism relies on Proactiv, which orchestrates the execution of workflows. Processes are embedded in virtual content via Docker's portable technology.

4. INTERCONNECTION APPROACH

The objective of the inter-connection is to allow CTEP users to launch specific processing services on the PEPS platform, using WPS requests. This requires knowing the available processing services and datasets, being able to request a processing, exchanging the results, and updating eventual accounting and catalogue information.

Processing service discovery and requesting is handled via the WPS interface. The WPS request is forwarded by the CTEP to the PEPS. Additional information not supported by WPS is included in the header of the request:

- User ID
- Job name: this will be used for the creation of the output result folder

After completion of the task, the output results must be sent to the user basket on the CTEP. For this, a “push” strategy has been adopted. A WebDAV interface allows to create a job output folder and copy the files. Following CTEP conventions, the output folder name is based on the job name defined by the user and a date.

5. USE CASE: A CHANGE DETECTION SERVICE

The interconnection is used to support a change detection service for the monitoring of protected coastal areas. The processor developed by Planetek (Italy) detects sudden decreases of the normalized vegetation index (NDVI) which can indicate an on-going illegal construction activity. An example of detection is shown in Figure 3.

This processor will be applied systematically in data driven mode over selected areas of interest in Sardinia, and will be tested manually in other areas. The service will rely on Sentinel-2 level 2 data, taking advantage of the 10 m resolution and 5 days revisit time.

If the experimentation is successful, the satellite-based monitoring service could become operational for environment protection authorities.



Figure 3: Example of new construction detected with the change detection software.

6. CONCLUSIONS AND PERSPECTIVES

Interconnecting several exploitation platforms offers the possibility to use seamlessly different data archives from a single access portal. The WPS standard simplifies somewhat this interconnection, but it does not support “administrative” aspects of the interconnection, such as user identification, authorization, and accounting. The standard was in fact designed to support client-server interactions rather than interconnections between processing platforms. In addition, there is no mechanism to transfer large processing results asynchronously.

In this experiment, we have used some add-ons to achieve these goals (additional headers, use of WebDAV). The WPS standard is sufficiently flexible to allow these extensions, but the solution is not completely satisfactory. An extension of the standard to support interoperability of platforms could be useful in the future.

7. REFERENCES

- [1] G. Sawyer, “Copernicus & Big Data: A Perspective from the European EO Services Industry,” Copernicus Big Data workshop, March 2014.
- [2] A. Annoni, “Copernicus and Big Data: Challenges and Opportunities”, Copernicus Big Data workshop, March 2014.
- [3] J. de La Mar, “Taking the data to the users - advanced networks and future Evolution”, Copernicus Big Data workshop, March 2014.
- [4] S. Loekken, J. Farres, “ESA Earth Observation Big Data R&D Past, Present, & Future Activities”, CEOS WGISS Meeting #37 Cocoa Beach, USA, 2014.

[5] E. Mondon, “Bringing the users to the data (+ applications and services)”, Copernicus Big Data workshop, March 2014.

[6] N. Gilles et al. , “The Coastal TEP: A Virtual Research Center for Coastal Environment Monitoring”, EO Open Science 2.0 Conference, Frascati, 2015.

[7] H. Jeanjean, “How to meet access requirements for the deluge of Sentinel data: the Sentinel Products Exploitation Platform (PEPS)”, Copernicus Big Data workshop, March 2014.

RF SIGNAL CHARACTERIZATION USING DEEP LEARNING

Ahmad Berjaoui, Adrien Elfassi

{ahmad.berjaoui,adrien.elfassi}@airbus.com
Datalab, Airbus Defence & Space

ABSTRACT

A deep learning model for automatic modulation recognition (*modrec*) has been developed. Two other models were developed: one for signal start and stop time estimation and another for residual frequency estimation. Modrec is performed with more than 90% accuracy in various noise and Doppler effect parameters, using a convolutional neural network mixed with recurrent units. Similarly, frequency and time estimation are very precise, reaching a small fraction of sampling frequency and symbol duration *resp.* These models have been trained and tested on simulated data, compatible with a typical space system with a ground emitter and a LEO receiver satellite.

Index Terms— Deep learning, modulation recognition, software radio, cognitive radio, dynamic spectrum access

1. INTRODUCTION

Deep learning has extensive applications in image processing, natural language processing and text processing. Deep learning for RF signal characterization is somewhat new as there are quite a few papers on the matter in literature [1], [2], [3], [4].

Applying deep learning to RF signal processing becomes natural after seeing numerous deep learning networks excel at distinguishing examples from the CIFAR-10 datasets [5], in image processing. Building on that example, one can naturally think of modulation recognition. Moreover, one of the main challenges in RF signal processing is synchronizing received signals in time and frequency. As deep learning models have also proven to be quite efficient for regression, it would also seem natural to use it to estimate propagation induced channel effects, such as Doppler frequency, Doppler rate and delay. Other effects such as multi-path fading and channel distortion could also be considered but these have been deemed secondary.

Moreover, in the context of a LEO satellite receiving multiple ground emissions, multi-path fading is no longer an issue.

The ability to recognize signal characteristics without going through a classical demodulation chain, opens a new field of applications, ranging from selective demodulation and collision detection, to dynamic spectrum access networking [2].

Special thanks to the Airbus Datalab

Typical space-based IoT applications [8] could greatly benefit from technical barriers that would be lifted by deep learning algorithms.

In section 2, we describe the considered signal model and pertaining physical effects. In section 3, training and test signal database generation parameters, are explained. In section 4, we describe the various models that were used to perform modrec and the achieved results. Finally, in section 5, we describe the various models that were used to estimate synchronization time and frequency, and the achieved results.

2. SIGNAL MODELING

2.1. Mathematical representation

All considered signals are assumed to be in baseband, and will be represented as follows:

$$s(t) = \sqrt{\frac{C}{N_0}} s(t - t_d) e^{j[2\pi((f_{res} + \delta_r t)t) + \phi(t)]} + \sigma(t) \quad (1)$$

where:

- $s(t - t_d)$ is the transmitted symbol signal and t_d is the propagation delay.
- $\sigma(t)$: normalized averaged white Gaussian noise (AWGN).
- f_{res} : residual signal frequency after down conversion. It includes the Doppler shift at receiver level.
- δ_r : the Doppler rate at receiver level. It cannot be neglected for low data rates, as can be typically found in IoT systems[8].
- $\phi(t)$: signal phase (including phase noise)

Time t is a multiple of the sampling frequency f_s , set at 100Hz. All other parameters will be computed accordingly, to be compatible with typical applications. This relatively small value helps reducing the amount of data that pass through the deep learning model.

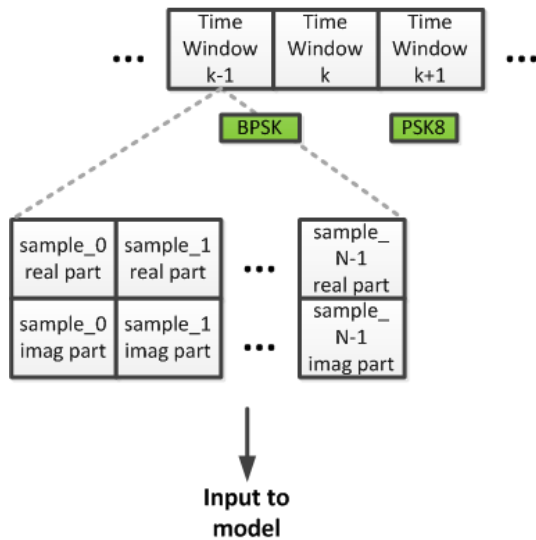


Fig. 1. Input samples windows: real and imaginary parts are stacked

2.2. Data format for deep learning

It is customary to consider RF signals as a stream of data, after the analog-to-digital conversion stage. However, working directly on a stream of data using software tools such as GNU Radio is not adapted in this case, as data labeling would be too complex. It is simpler to work on time-windows of fixed length, i.e on the same number of samples, for the training process. Once the model has been properly trained, inference can be inserted into a streamed processing chain by slicing the data stream into similar time-windows. However, for this to work properly, the training set must include cases where only a portion of the useful signal is present, i.e models would have to be able to deal with only a segment of message or signal frame.

Most deep learning development libraries make use of real valued functions. The typical complex signal representation is not suitable and can be easily replaced by stacking the real and imaginary parts of the signal. Each signal becomes an image of height 2 and width the signal’s length, as shown in fig.1.

2.3. Deep learning modeling strategy

The main advantage of working with a fixed number of samples is that we inherently transform the modrec problem into a typical “image” classifier scenario, where image labels have been replaced by digital modulation types. In this context, convolutional networks such as in [1] or [6] should yield excellent accuracy.

Working on a data stream is not necessarily the best strategy but we cannot ignore the sequential nature of data. Therefore, networks that employ recurrent units such as LSTM [7] and GRU[10] were also tested, and the best overall performance was achieved by combining both types of networks.

Dataset type	Simple	Intermediate	Challenging
SNR [dB]	5 20	5 20	5 20
Doppler shift [Hz]	0	0	-5 5
Doppler rate [Hz/s]	0	0	-1 0
Phase offset [°]	0	-20 20	-20 20

Table 1. Datasets generation parameters

3. TRAINING & TEST DATASETS

Seven types of signals were considered: AWGN, CW, BPSK, QPSK, 8PSK, 16QAM and 64QAM. For each type of signal, 120 000 examples were generated, thus creating a perfectly balanced dataset. Several datasets were generated, ranging from high SNR and no signal disturbance (simple scenario), to low SNR and all types of signal disturbances (challenging scenario). In all cases, the generated signal can occur at any time within the window frame. Generation constraints were added to insure that a minimum number of samples are present in the generated window. The Doppler shift (f_{res}) and rate (δ_r) correspond to what could be expected in a LEO satellite receiver scenario[8], considering the 100 Hz sampling frequency. Phase offset ($\delta\phi$) is randomly drawn (uniformly) and is assumed constant for a given message. The generation parameters are summarized in table1.

The impact of the number of samples per window was also studied. 128 and 512 were considered and datasets of each type were generated accordingly, using a Spark cluster of over 150 vCPUs to reduce computation time (a few minutes instead of 1.5 hours per dataset).

4. MODULATION RECOGNITION

The first tested model is the one described in [1]. It uses windows of 128 samples as inputs. It was retrained using our datasets and achieved a score (accuracy) between 65% and 80% (challenging and simple scenario *resp.*). A modified version, allowing it to handle windows of 512 samples yielded a score between 69% and 82%. As can be naturally expected, increasing the number of samples by window yields better results.

Inspired by the efficiency of the Inception architecture [6], which consists in concatenating the outputs of different types of convolutions between each layer, we have tested a model with 3 inception layers, followed by 2 fully connected layers. Results were far better for the simple dataset (up to 96% accuracy using 512 samples) but still low (70.4%) in the challenging scenario.

The best results were achieved by using 512 samples win-

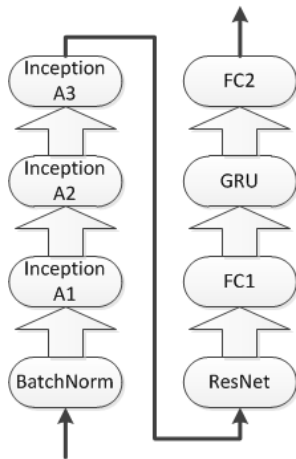


Fig. 2. Modrec Inception-ResNet-GRU model

	AW GN	B PSK	Q PSK	8 PSK	16 QAM	64 QAM	CW
AW GN	100	0	0	0	0	0	0
B PSK	0	98.8	0.4	0.6	0.1	0.1	0
Q PSK	0	0.1	88.7	11.2	0	0	0
8 PSK	0	0	8.6	91.3	0	0.1	0
16 QAM	0	0	0	0	88.7	11.2	0.1
64 QAM	0	0	0.1	0	15.0	84.9	0
CW	0	0	0	0	0	0	100

Table 2. Modulation recognition confusion matrix for the best model over the most challenging dataset. The overall average accuracy is 93.2%

dows, a model composed of inception layers, GRU cells and a residual structure[6]. This model consists of 3 consecutive inception layers, an inception ResNet layer, a fully connected layer, a GRU grid of 256 cells and a fully connected layer. Models were trained using a 32 vCPUs/Tesla K80 GPU virtual machine.

The detailed confusion matrix using windows of 512 samples is provided in table2.

As can be naturally expected, table.2 clearly shows how hard it is to distinguish close high-order phase modulations such as QPSK/8PSK and 16QAM/64QAM *resp.* All tested models have shown little difficulty distinguishing AWGN and CW from phase modulated signals.

5. TIME & FREQUENCY SYNCHRONIZATION

RF characterization of a digital message also requires precise time and frequency synchronization, especially for sub-

sequent demodulation. As the beginning and the end of a message are unknown, one must estimate both the beginning of the message (t_{start}) and the end of the message (t_{end}). The same applies to the message's residual frequency. Since a Doppler rate is not excluded, the residual frequency at the beginning of the message ($f_{start} = f_{res}$) may differ than the one at its end ($f_{end} = \delta_r t_{end} + f_{start}$). Time/frequency prediction accuracies are expressed as follows:

$$\begin{aligned} start_{acc} &= |t_{start} - \hat{t}_{start}| \\ end_{acc} &= |t_{end} - \hat{t}_{end}| \end{aligned} \quad (2)$$

where \hat{t}_{start} , \hat{t}_{end} , refer to the start/end time estimations *resp.* The same equation applies to start/end frequencies.

The estimators are fed not only with the raw samples but also with the modulation type. Indeed, in a full blind RF characterization chain, modrec would have been performed as a first step, and the obtained information could ease the subsequent synchronization step. Tests were performed with all the aforementioned modulations (except AWGN and CW). For the sake of simplicity, we have decided to focus on the 16QAM waveform, which offers a good compromise between spectral efficiency and synchronization complexity. We have also considered only datasets with 128 samples, making it harder to synchronize.

The estimators' performances are measured over the accuracies' cumulative distribution at 90%. The Cramer-Rao lower bounds (CRLB) for time and frequency estimators can be computed using eq.3[9]. In fact, $CRLB(\hat{f})$ is the lower bound for Doppler shift estimation, assuming it is the only contributor to the residual frequency :

$$\begin{aligned} CRLB(\hat{t}) &= \frac{0.55}{\sqrt{2B\sqrt{B.T.SNR}}} \\ CRLB(\hat{f}) &= \frac{0.55}{\sqrt{2T\sqrt{B.T.SNR}}} \end{aligned} \quad (3)$$

where B is the rectangular spectral band of interest (40 Hz, considering Nyquist with some margin) and T is the RMS message time. Assuming 128 samples 16QAM messages with 10% of symbols being used for time synchronization (e.g. preamble symbols), $T = 0.1 \cdot N_{symb} \cdot \log_2(16) / f_s$. Numerically, this yields (at $SNR = 20dB$):

$$\begin{aligned} 1.64 * CRLB(\hat{t}) &= 1.2ms \\ 1.64 * CRLB(\hat{f}) &= 0.96Hz \end{aligned}$$

The developed estimators provide an estimation for the "start" and "end" time and frequency *resp.* As no state-of-the-art model was found, we have tested several CNN with gradual complexity, both for time and frequency. The best results were achieved using Inception ResNet structures, as shown in fig.3. The same model structure is used for time and frequency. Features are first extracted independently for "start"

	Start time (ms)	End time (ms)	Start frequency (Hz)	End frequency (Hz)
Accuracy	93	44	3.9	4.4

Table 3. Time and frequency estimators accuracy using the "challenging" data set (16QAM windows only) at 20 dB of SNR

and "end". These features are fed directly to the final fully connected layers ("FC A" for "start" and "FC B" for "end"). Common features are extracted by the 2x2 convolution layer, and then concatenated with the modulation type. The resulting features are then fed into the final fully connected layers. The estimators' performances are given in table.3

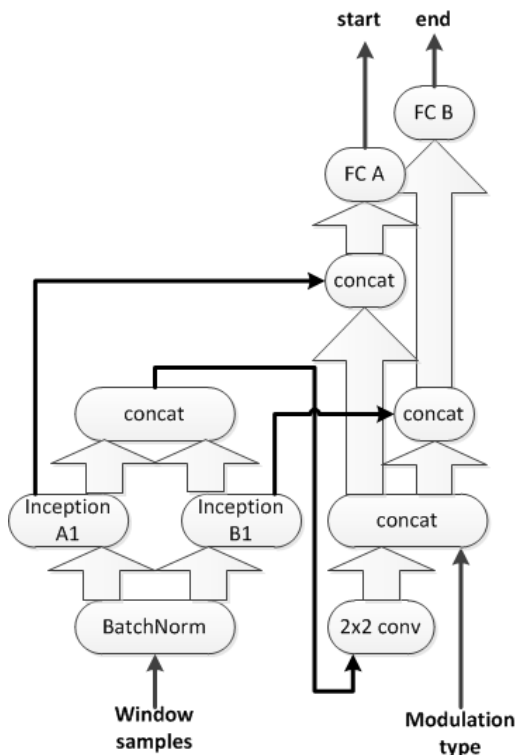


Fig. 3. Start/end time & Frequency estimation model

6. CONCLUSION & FUTURE WORK

We have developed a modulation recognition deep learning classifier, capable of distinguishing 7 types of modulations (with an average accuracy of 93.2%), in challenging conditions, representative of a LEO satellite receiver configuration. We have also developed a deep learning regressor (with a very good accuracy w.r.t the CRLB), capable of estimating time and frequency synchronization parameters of blind messages. These models can still be improved in terms of accuracy but also simplified (e.g. with thorough hyper-parameter tuning). They can also be trained to detect colliding mes-

sages, and with some further development, they could separate them. RF characterization is an essential step in dynamic spectrum access[1],[2],[3] and could relieve some of the constraints of harsh frequency regulations, especially in space-based telecommunication systems.

7. REFERENCES

- [1] Timothy J. O’Shea et al. *Convolutional Radio Modulation Recognition Networks*. Engineering Applications of Neural Networks: 17th International Conference, 2016.
- [2] Nathan E West, Timothy J O’Shea. *Deep Architectures for Modulation Recognition*. IEEE International Symposium on Dynamic Spectrum Access Networks, 2017.
- [3] O’Shea et al., & Clancy, T. C. (2016, November). *Radio transformer networks: Attention models for learning to synchronize in wireless systems*. In Signals, Systems and Computers, 2016 50th Asilomar Conference on (pp. 662-666). IEEE.
- [4] O’Shea, T. J., Karra, K., & Clancy, T. C. (2016, December). *Learning to communicate: Channel auto-encoders, domain specific regularizers, and attention*. In Signal Processing and Information Technology (IS-SPIIT), 2016 IEEE International Symposium on (pp. 223-228). IEEE.
- [5] Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). *Deep learning*. Nature, 521, 436-444.
- [6] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. In AAAI (pp. 4278-4284).
- [7] Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*. Neural computation, 9(8), 1735-1780.
- [8] Anteur, M., Deslandes, V., Thomas, N., & Beylot, A. L. (2015, December). *Ultra narrow band technique for low power wide area communications*. In Global Communications Conference (GLOBECOM), 2015 IEEE (pp. 1-6). IEEE.
- [9] S. Stein, (June 1981). *Algorithms for Ambiguity Function Processing*. In IEEE Trans. On ASSP, June 1981.
- [10] Chung, Junyoung, et al. *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv:1412.3555 (2014).

OFF THE SHELF DEEP LEARNING PIPELINE FOR REMOTE SENSING APPLICATIONS

Rachit TRIPATHI¹, Adrien CHAN-HON-TONG² and Alexandre BOULCH²

1: QuantCube Technology

2: ONERA, the french aerospace lab

ABSTRACT

Designing specific index for a some remote sensing applications require a large research effort not scalable to the multitude of applications.

Inversely, using off the shelf deep learning pipeline could be good enough for some applications.

We describe off the shelf deep learning application on the 2017 data fusion contest (IEEE-IGARSS) for local climate zone estimation. While being completely non expert to local climate zone estimation, and while having only few meta parameters, these pipelines reach honorable scores on this dataset compared to hard to tune winner pipeline of the challenge.

Index Terms— deep learning, remote sensing

1. INTRODUCTION

The popularization of remote sensing images (e.g. the free availability of sentinel images) could allow a rupture for large remote sensing applications including climate observation, biomass estimation, drought monitoring... However, seeing the large spectrum of possible applications of remote sensing images, we can wonder about the research effort to correctly extract information from these new data. Designing specific index to handle specific problem may not be an scalable way to take full advantages of all these newly available data.

Inversely, we claim that using off the shelf deep learning pipeline could be good enough for some applications.

To argue our statement, we present here experiments done on the data fusion contest 2017 (IEEE-IGARSS). Data fusion contest (DFC) are a set of challenges of the remote sensing community. The 2017 challenge is about predicting Local Climate Zones (LCZ) from training cities to unknown cities [1]. LCZ aims to offer a typology of both landscape and urban locations designed to study heat island and heat propagation in cities. The LCZ of a location is completely defined by it landscape/urban configuration e.g. high dense metallic building will lead to LCZ named *dense high rise*.

We show that off the shelf deep learning pipelines can reach honorable scores for LCZ estimation (compare to the state of the art) without requiring careful tuning.

In the context of the DFC2017, provided inputs for doing this LCZ prediction are remote sensing images and crow

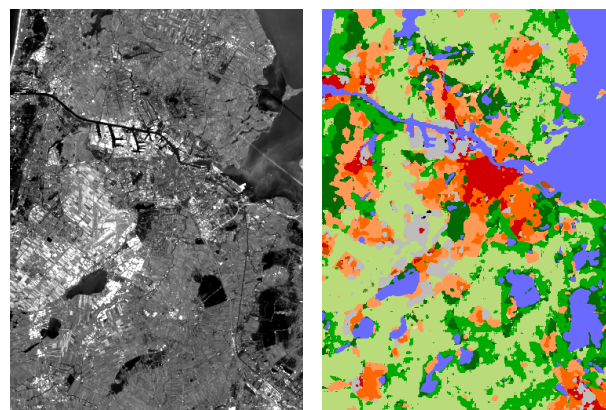


Fig. 1. Landsat 8 data (band 4) and predicted LCZ map on the test city of Amsterdam.

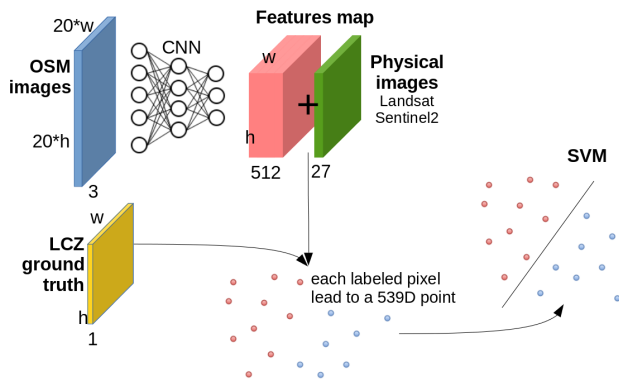
based map: 9 bands *landsat* at a resolution of 100m with multiple images per city, 9 bands *sentinel2* images at 100m, *openstreetmap* information available for the selected cities (rasterized at 5m) and not registered 9 bands *sentinel2* images at 5m.

Currently, even before the data fusion contest, there were works on LCZ estimation from public data (e.g [2]). With the challenge, there were a large research effort on this problem [3, 4, 5, 6]. In [3], only *landsat* and *osm* are handled. Both basic and specific features are extracted from images. Basic features are mean-variance. Specific ones are mostly built on infrared measure. Thus, a first step of atmospheric correction is performed on *landsat* images to improve infrared images. Then well known normalized difference index (we will call it *ndi*) like NDVI, NDWI, MNDWI, NDBI, BSI and WRI (see [7] for a brief review) are extracted from images. In addition, morphological profiles are extracted by combining *osm*, NDVI and morphologic operator. Then, two kind of ensemble classifier are trained on these features.

Seeing [2, 3] ([4, 5, 6] are currently not available), we can argue that most of these LCZ papers either needs careful tuning or use very specific index designed especially for LCZ. Here, we offer instead very generic off the shelf deep learning pipelines to infer LCZ from multi modal data.

Table 1. Results of the leave one city out

method	berlin	HK	paris	rome	sao paulo	average
images (1vsall)	42%	25%	74%	22%	43%	38%
osm (1vsall)	47%	43%	59%	33%	21%	36%
images + osm (1vsall)	53%	51%	72%	30%	54%	48%
cnn + svm (raw + osm)	50%	52%	73%	33%	68%	51%
images + ndi + osm	51%	53%	73%	34%	68%	52%
cnn + svm (ndi + osm)	57%	53%	67%	48%	52%	54%
CNN Pyramid Pooling	71%	67%	69%	51%	80 %	67.6%

**Fig. 2.** scheme of the cnn+svm pipeline.

2. DEEP LEARNING FOR LCZ

2.1. CNN + SVM

The first pipeline is inspired from [8]. It is simply built by extracting a set of feature maps directly or with convolutional deep neural network (CNN) and using support vector machine (SVM) to perform pixelwise classification (precisely libsvm with 1vs1, linear kernel and default parameter [9]). Notice that this pipeline has 0 parameter, and thus, is a good example of off the shelf pipeline (see figure 2). Code is available here https://github.com/achanhon/CNN_SVM_for_DFC2017.

Raw images are used directly as features. For landsat data, we compute the mean and variance maps for each channel, in such a way that we exploit the multiple acquisitions. Then, we concatenate all provided bands for both landsat and sentinel leading to a 27 image based features per 100m pixel.

Additional features are generated using VGG16 [10] initialized imagenet weights on OSM data. We compute an ad hoc mask per city from osm data: it is formed using building, green area from landuse, and road (pixel value is either 0 or 255 depending on the presence of the item in osm). The features are extracted at several layers and rescaled to 100m resolution. In our experiment, osm typically provides 512 features per 100m pixel. Notice that, we do not train the cnn from an imagenet initialization, instead we just use the pretrained version without adjustment in the convolution weight (only small change have been done on the pooling structure to take

into account the difference in resolution between images and osm).

2.2. Late and cascaded fusion with CNN

We used OSM and sentinel2. We design a specific networks inspired from U-Net [11] for each data sources, each with more pooling than upsampling as data are more resolved than label maps. Then, we concatenate resulting maps and forward it in a second network.

In late fusion, this second network is just a series of convolutions (since convolutions does not reduce a size of feature maps, we can use them directly to predict the labels). It is a *late* fusion because each modality is processed independently and only high level representation are combined at the top of the network.

We also evaluated cascaded fusion in which the second network is large. Two architecture have been tested for this second network.

1. U-Net (again): we concatenate upsampled lower layer with top ones to predict the labels.
2. Pyramid Pooling like: Inspired from [12], we implement a network with different levels of pooling along with upscaling and concatenation plus a final convolution to get the prediction.

2.3. Early fusion with CNN

The final approach uses all the data available in the challenge with additional Sentinel 1 data. The Sentinel 1 composite (S1 composite) is a three channel composite: VV polarization for ascending acquisition, VH for ascending and descending and VV for descending. The final product is the mean over the year 2016.

The CNN architecture used in this section is based on SegNet [13]. We set up an encoder for each input type (Sentinel 1 and 2, Landsat 8 and OSM). The decoder input is the concatenation of all coded signals. The reference signal (the one use for unpooling operation) is Landsat 8. Compared to the late fusion (previous section), we merge features instead of activations that why we can speak of early fusion. See figure 3 for visual representation of the 3 different CNN pipelines.

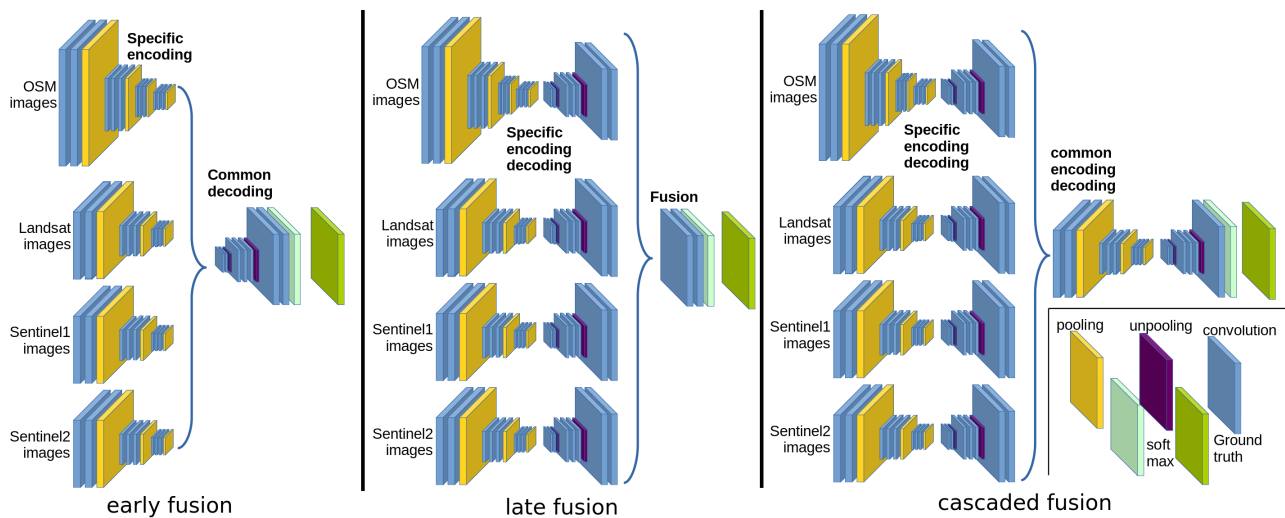


Fig. 3. Visualization of the 3 different cnn pipelines.

This is a schematic visualization and not the real architectures of the 3 pipelines. In all these 3 pipelines, weight are optimized by stochastic gradient descent (and not just restored from pretrained model). Training is done in classic fashion: forwarding the input, we get an output which has the same shape as expected ground truth ; a loss measures the distance between the produced output and the ground truth resulting in a gradient which is computed by backpropagated across the network and used to update all weights.

Table 2. Results on the test database

method	test
cnn + svm (raw + osm)	58%
cnn + svm (ndi + osm)	57%
fusion Pyramid Pool.	52%
Early fusion	56.6 %
Early fusion additional cities	64.3 %

We trained two models: one on the cities of the DFC2017 training set and, in order to estimate the influence of the training data set size, one with additional cities: Dublin, Houston, Sydney, Vancouver and Warsaw. The ground truth associated with these cities is denser but with poor quality: coarse areas and noised labels.

3. EXPERIMENT

Following the rules of the 2017 data fusion contest, we evaluate the quality of the LCZ estimation by measuring the pixel wise accuracy (this is so a semantic segmentation problem).

The server evaluation results are presented in table 2 (notice that inspired from [3], we both use raw image and ndi in cnn+svm).

In addition, to the server evaluation, we also provide a leave one city out protocol: all training cities except one are used to train the model which is then applied to the excluded city, this operation being done for all cities. Leave one city out

results (when available) are detailed in table 1 with an *average* result. In order to penalize, very unstable results across cities we weight the worse accuracy by a factor 2 in the *average* result.

4. DISCUSSION

The main result is that these off the shelf pipelines compare honorably, in our opinion, to hard to tune state of the art of LCZ estimation which is between 69 to 74% (see the output of [1]).

Indeed, due to small size of the ground, heavy CNN models like late or cascaded fusion strongly overfits and thus are overcome by models consisting in training only the SVM on the test set. This is consistent with the idea that deep learning is mainly relevant when large amount of data/ground truth is available. However, even in context with small amount of data, using pretrained cnn can provide interesting result (cnn+svm still achieves 58%). And, the required size of the dataset may not be that high: in our experiment, additional training data consisting in only 4 new cities gives a real boost in performance while being corrupted training data. Thus, we can be not so worry about preventing the model from overfitting.

Finally, we notice that the osm and image data are very complementary: both images and osm alone reach low performance probably because some classes are not distinguishable using only one modality (water is not present in osm resulting in a complete impossibility to predict at least this classe) and

image may not be sufficiently resolved to infer density and thus elevation of building without osm. By the way, it is still not clear if image+osm are sufficient to distinguish between some classes like *middle rise* and *low rise*.

5. CONCLUSION

In our opinion, the main result of this work is that our off the shelf pipeline reach honorable results seeing the state of the art without requiring large tuning.

Off course (may be hopefully) designed index and algorithm performs still better than off the shelf deep learning (at least in our experiment ndi largely increases performance stability over different cities). But, this example highlights the interest of off the shelf deep learning pipeline to take advantage of newly available remote sensing image.

Thus, we argue that off the shelf deep learning pipelines may be more and more present for remote sensing applications.

Acknowledgment

As user of the DFC2017 data, the authors would like to thank the WUDAPT (<http://www.wudapt.org/>) and GeoWIKI (<http://geo-wiki.org/>) initiatives for providing the data packages used in this study, the DASE benchmarking platform (<http://dase.ticinemaerospace.com/>), and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. Landsat 8 data available from the U.S. Geological Survey (<https://www.usgs.gov/>). OpenStreetMap Data OpenStreetMap contributors, available under the Open Database Licence <http://www.openstreetmap.org/copyright>. Original Copernicus Sentinel Data 2016 available from the European Space Agency (<https://sentinel.esa.int>).

6. REFERENCES

- [1] D. Tuia, G. Moser, B. Le Saux, B. Bechtel, and L. See, "2017 ieee grss data fusion contest: Open data for global multimodal land use classification [technical committees]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 70–73, March 2017.
- [2] P. Lopes, C. Fonte, L. See, and B. Bechtel, "Using openstreetmap data to assist in the creation of lcz maps," in *2017 Joint Urban Remote Sensing Event (JURSE)*, March 2017, pp. 1–4.
- [3] Naoto Yokoya, Pedram Ghamisi, and Junshi Xia, "The data fusion contest 2017: Open data for global multimodal land use classification," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.
- [4] S. Sukhanov, I. Tankoyeu, J. Louradour, R. Heremans, D. Trofimova, and C. Debes, "Multilevel ensembling for local climate zones classification," 2017, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.
- [5] Camila Souza dos Anjos, Marielcio Goncalves Lacerda, Leidiane do Livramento Andrade, and Roberto Neves Salles, "Classification of urban environments using feature extraction and random forest," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.
- [6] Yong Xu, Fan Ma, Deyu Meng, Chao Ren, and Yee Leung, "A co-training approach to the classification of local climate zones with multi-source data," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017.
- [7] Komeil Rokni, Anuar Ahmad, Ali Selamat, and Sharifeh Hazini, "Water feature extraction and change detection using multitemporal landsat imagery," in *Remote Sensing*, 2014.
- [8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014.
- [9] T. Joachims, T. Finley, and Chun-Nam J. Yu, "Cutting-plane training of structural svms," in *Machine Learning*, 2009.
- [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *CoRR*, 2014.
- [11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Springer International Publishing, 2015.
- [12] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [13] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla, "Segnet a deep convolutional encoder decoder architecture for robust semantic pixelwise labelling," in *arXiv preprint*, 2015.

DEEP LEARNING FOR DENOISING OF SATELLITE IMAGES

Pierre Blanc-Paques¹, Renaud Fraisse

Datalab & Image Chain department, Airbus Defence and Space, Toulouse

Abstract

The quality of satellite images is of primary importance to accurately identify and extract valuable information in the data. While the image is affected by different sources of noise, the goal of denoising is to recover the initial signal without loss of information. Current state of the art in image denoising is achieved by advanced mathematical methods (non-local filtering, wavelet transforms), yet recent advances in deep learning have brought up new interesting approaches. We explore these approaches and show that, by leveraging efficient tools from the big data world, the deep learning can be attractive for satellite imagery: both in terms of performance and industrialization process.

Index Terms— Satellite imagery, denoising

1. INTRODUCTION

The image-quality requirements for the next generation of very high resolution satellites are particularly stringent. For the denoising, the goal is to improve the SNR (Signal to Noise Ratio) by more than 40%. Considering the huge amount of data downloaded by the satellites, the processing also has to be extremely efficient: a panchromatic image of 30000 * 30000 pixels shall be processed in less than a few minutes.

State of the art methods like NL-Bayes or BLS-GSM [1], [2] achieve performances which are in line with the image quality requirements. However these algorithms need an expert hand to be finely tuned and their implementation has to be optimized to cope with the computation-time requirements. Besides, it is worth exploring new ways of further improving the denoising performance, as any gain on this side could reduce the constraints on the instrument and result in massive cost savings

In terms of workflow, deep learning techniques offer an interesting approach for the denoising. First of all, the essence of deep-learning is to train several versions of the algorithm and to select the best one. Efficient tools and platforms [3] are available to perform this selection, thus greatly reducing the need for experts to handcraft the best possible tuning. In addition, powerful GPU implementations exist for the convolutional neural networks [4], making the

run-time performances of large convolutional models extremely competitive compared to more traditional approaches. One of the main question is thus to determine whether deep learning algorithms can achieve the image quality requirements for denoising.

2. DEEP LEARNING FOR IMAGE DENOISING

2.1. Models

The principle of deep learning for denoising is to train a deep neural network for predicting the noise-free image from the input noisy image. A classical approach is to use an autoencoder [5], [6]. The autoencoder has several hidden layers that can be seen as a code used to represent the input: the goal of the autoencoder is thus to learn robust high-level representation of the data by taking the noisy image as input and attempting to fully reconstruct the reference image.

Another recent approach is the residual network. Instead of directly computing the noise-free image, the network outputs the difference between the noisy observation and the target noise-free image. As the residual network learns a function which is very close to the identity, it is much easier to train and produces better results [7]. In this paper we focus on residual networks with architecture similar to the one described in Figure 1: we use Tensorflow library to define networks stacking from 6 to 12 {convolution – batch normalization – Rectified Linear Units (ReLU)} layers with one shortcut connection. It is sufficient to limit the number of layers to 12 because it appears that at some point, stacking more layers brings marginal gain to the performance while increasing the computation cost. On this point, our architecture slightly differs from the one proposed in [7], where up to 20 layers are stacked: this is probably due to the lower noise level that we have to deal with. During the training, the cost function is the L2 norm between the reference noise free image and the output of the network.

2.2. Training data

The training and evaluation dataset is built by simulating the full image chain: the image acquisition process is modelled by applying the instrument blur to a perfect reference image, and finally applying the noise.

¹ computervision@airbus.com

3. PERFORMANCES EVALUATION

Several algorithms have been compared on representative images and scenes. The results are proposed on Table 1. Based purely on these metrics, it is clear that the residual network (ResNet) offers denoising performances which are comparable to state-of-the-art algorithms (NL-Bayes and BLS-GSM).

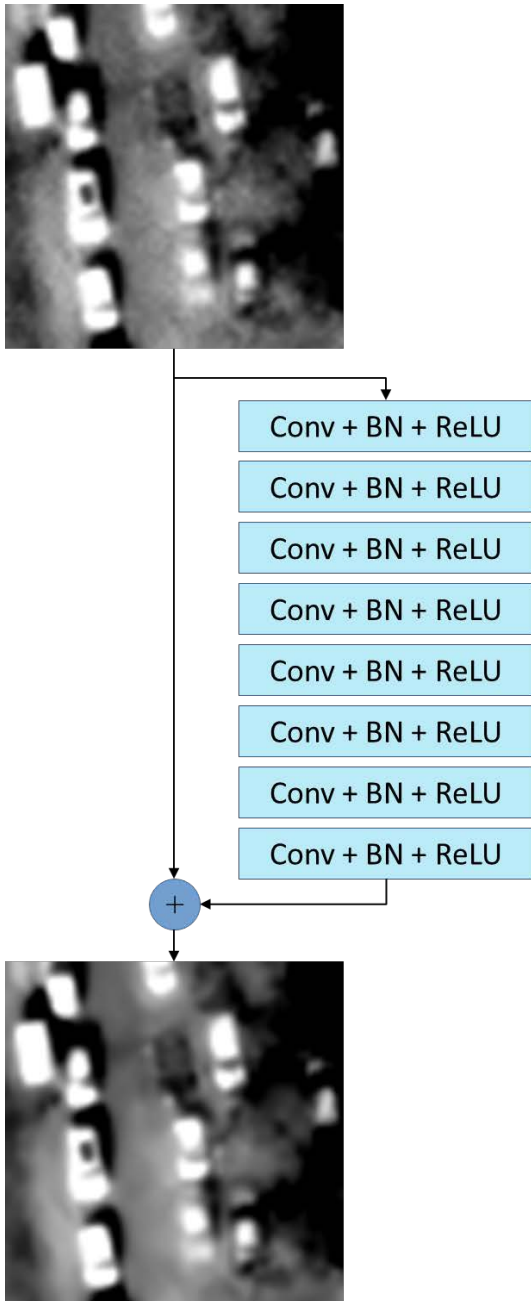


Figure 1: Deep residual network for denoising

Table 1: Image quality metrics for different algorithms²

	RMSE (LSB)	Max Error (LSB)	99.7% error (LSB)	RMSE local (LSB)	PSNR	SSIM
No denoising	6.5	43	21	6.4	45.2	0.90
NL-Bayes	4.0	34	14	3.7	49.6	0.97
BLS-GSM	4.1	31	14	3.9	49.3	0.97
ResNet	4.0	35	14	3.8	49.4	0.97

Please note, however, that additional qualitative image analysis can be necessary to rank the algorithms based on the application. This point is particularly interesting, as it highlights the fact that the traditional metrics (RMSE, PSNR, SSIM) are not sufficient to fully capture the operators requirements for the image quality.

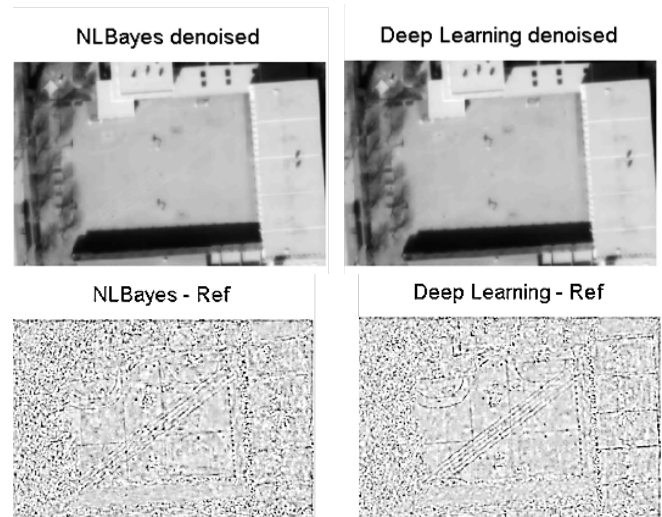


Figure 2: Denoising by NL-Bayes and Deep Residual network

On this particular aspect, one of the difficulties with the deep network is that it tends to remove high-frequency events, which is penalizing for human interpreters (example on Figure 2: the thin straight lines in the middle of the playground; here the bottom row shows the difference between the output image and the reference noise free image). This is most likely due to the L2-norm used as the cost function for training the network: it has indeed been shown [8] that for image generation problems, the L2-norm somehow discards high-frequency contents and produces blurry images.

² Metrics: Root Mean Square Error (RMSE), maximum error, error at 99.7%, local RMSE, Peak SNR (PSNR), Structural SIMilarity (SSIM).

Several approaches are currently envisioned to define metrics more adapted to the interpreters' requirements: errors with penalty factors on high-frequencies, content loss [9] (distance between feature representations of reference image and output image, where the features are obtained by a dedicated neural network), generative networks. A thorough analysis of these approaches will have to be undertaken in a future work.

In terms of run-time performances, the deep residual network benefits from the powerful GPU implementations of convolutional networks, which is not the case for classical algorithms. This leaves the deep learning approach unchallenged in terms of run-time, with a processing time of 20s for a 3000 * 3000 image on one GPU (time decreases linearly with the number of GPU used) versus several minutes for other algorithms (C++ CPU implementation).

From an industrial point of view, another attractive aspect of deep learning is the development and integration process: once the training is performed, the resulting model can be directly encapsulated in a docker container and exposed as a service in the ground segment infrastructure. In the traditional V-cycle, the algorithm design is performed based on the functional requirements, then the software requirements are derived, and then the software is developed, validated, integrated, and finally qualified. Here the approach is to go straight from design to integration by directly reusing the code generated during the training (which can be seen as some sort of autocoding), potentially resulting in huge planning and cost savings.

On the industrial side, it is also interesting to note that the deep learning approach offers a robust framework for addressing different kind of problems. Once the process is setup, new efficient algorithms can be created and integrated in a very short time frame. For instance, Figure 3 shows the result of compression noise removal on the exact same setup as the one described earlier.

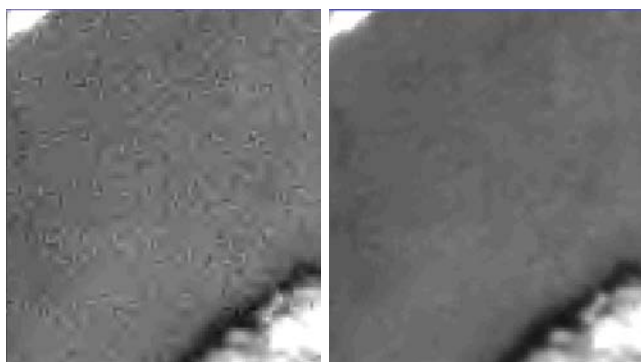


Figure 3: Example of compression noise removal

4. CONCLUSIONS AND FUTURE WORK

We have proposed an introductory analysis of deep learning for satellite image denoising. We have shown that deep residual networks achieve RMSE and PSNR comparable to state-of-the-art denoising algorithms, run faster, although the visual quality of the results is not as good. Yet, by leveraging powerful tools and infrastructures, deep-learning seems like a good way to quickly setup decent operational solutions. Future work will thus have to consolidate this analysis by addressing:

- The improvement of algorithm performances and the definition of metrics better adapted to capture the interpreters' image quality requirements.
- The robustness of the deep learning solution versus the representativeness of the models used in the simulation of the noisy images (construction of the training set).
- The interest of using deep learning for the full restoration chain (deconvolution, denoising).

5. REFERENCES

- [1] Lebrun, M, Buades, A, Morel, J-M, "Implementation of the NL-Bayes image denoising algorithm", *Image Processing On Line* 3, pp. 1-42, 2013
- [2] Rajaei, B, "An analysis and improvement of the BLS-GSM denoising method", *Image Processing On Line* 4, pp. 44-70, 2014
- [3] Google Research, "Tensorflow: large-scale machine learning on heterogeneous distributed systems", *White Paper*, 2015
- [4] Zeiler, D, Fergus, R, "Visualizing and understanding convolutional networks", *Arxiv*, 2013
- [5] Vincent, P, Larochelle, H, Bengio, J, Lajoie, I, Bengio, Y, Manzagol, P-A, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion", *Journal of Machine Learning Research* 11 pp. 3371-3408, 2010
- [6] Gondara; L, "Medical image denoising using convolutional denoising autoencoders", *Arxiv*, 2016
- [7] Zhang, K, Zuo, W, Chen, Y, Meng, D, Zhang, L "Beyond a Gaussian denoiser: residual learning of a deep CNN for image denoising", *Arxiv*, 2016
- [8] Larsen, A, Sonderby, K, Winther, O, "Autoencoding: beyond pixels using a learned similarity metric", *Arxiv*, 2015
- [9] Ledig, C, Theis, L, Huszar, F, Caballero, J, Cunningham, A, "Photo-realistic single image super-resolution using a generative adversarial network", *Arxiv*, 2016

DATA INTEGRATION OF REMOTE SENSING AND IN SITU DATA FROM SEVERAL SOLAR SPACE MISSIONS FOR SPACE WEATHER SERVICES

M. Casti^{1,2}, S. Fineschi², R. Messineo¹, E. Antonucci², A.F. Mulone¹, A. Bemporad², A. Fonti¹, R. Susino², F. Filippi¹, D. Telloni², F. Solitro¹, G. Nicolini², M. Martino¹

¹ALTEC S.p.A., Corso Marche 79, 10146 Torino, Italy

²INAF-OATo, Via Osservatorio 20, 10025 Pino Torinese, Torino, Italy

ABSTRACT

Remote sensing and in situ open data relative to the Sun, the heliosphere and the Earth's magnetosphere, combined with the novel big data technologies, give scientists the possibility to design, implement and validate space weather algorithms on extensive datasets. This paper introduces the Heliospheric Centre project for space weather medium-term and short-term forecast, which is under development in Turin (Italy), focusing on the technical part of data integration. Firstly, the data flow of the forecast pipeline is described, highlighting the constraints of real time delivery for forecast services. As second point, data integration is discussed in order to underline points such as data model conversion, metadata organization and data normalization. Then, the data management elements are pointed out to inform about the architecture and technologies used to implement the system. Finally, the future prospects of the Heliospheric Space Weather Centre are described.

Index Terms — Space Weather, Forecast, Big Data Management, Data Centre, Remote Sensing, In situ, Solar Corona, CME

1. INTRODUCTION

Modern society is characterized by a strong dependence on electrical/electronic and space-based technologies. Consequently, the vulnerability to space weather related phenomena is rapidly increased: a century ago, the main impact was associated to the telegraph system now the risks are related to a larger amount of technologies; e.g., electric power grids, satellites, navigation and timing information systems and long-range high-frequency radio communications. In order to shield the society against the potential damaging effects of the space weather, an improvement of its forecast is needed.

It is well known that space weather domain is vast - starting from the Sun until outside the planetary orbit - and that the related physical phenomena are complex. For these

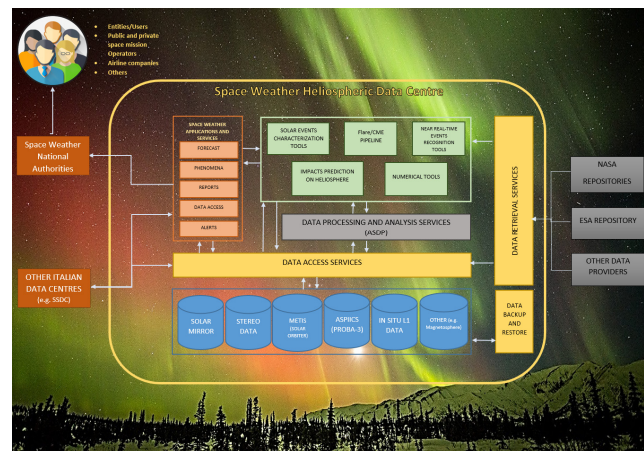


Figure 1: Space Weather Heliospheric Data Centre

reasons, as it is well illustrated in [1], an international effort is required in order to better understand the driving phenomena and to provide an efficient alert service to society.

This paper aims at describing the Heliospheric Space Weather Centre project, which combines scientific research with engineering inventiveness for space weather forecast.

2. HELIOSPHERIC SPACE WEATHER CENTRE PROJECT

The Heliospheric Data Centre is a joint effort between ALTEC and INAF-OATo, both located in Turin, Italy. The project has two main objectives.

The first one is to consolidate and evolve the Heliospheric Data Centre, initially set up with the SOHO data coming from the ESA approved SOLAR (SOHO Long-term ARchive) archive, in order to manage additional solar archives storing solar coronal and heliospheric data coming from ESA and NASA space programs.

The other one is to develop a Heliospheric Space Weather Centre for forecast of impacts of solar disturbances on the heliosphere and the Earth's magnetosphere.

2.1. Medium-term and short-term forecast services

According to the World Meteorology Organization (WMO), the meteorological forecast is classified in accordance with the time range considered for predictions. In our understanding, in space weather literature there is not a similar well-defined classification. For this reason, in this paper we will consider the following terminology: medium-term forecast indicates weather forecasting on a period less than ~ 2 days, while short-term forecast indicates a weather forecasting on a very short-term period of up to ~ 1 hour.

The medium-term and short-term forecast prediction services provided by the Heliospheric Space Weather Centre will consist of the identification of high-speed solar wind streams and solar ejections of magnetic clouds and of the alerting for those reaching the Earth magnetosphere. The medium-term forecast services are based on processing remote sensing observations of the Sun and the heliosphere; these services could predict up to few days before the arrival onto the Earth's magnetosphere of solar phenomena driving major geomagnetic storms. On the other hand, short-term forecast services process the in situ heliospheric observations at the Sun-Earth Lagrangian point L1, in order to predict geomagnetic storms up to a few hours before their occurrence.

2.2. Processing Pipeline

The Heliospheric Space Weather Centre processing pipeline is composed of two different branches that interact in order to achieve a medium and a short-term prediction of solar driven phenomena. The first branch is design to process remote sensing data and to provide the medium-term forecast. On the other side, in situ data are processed.

The first step of the processing pipeline is represented by the ingestion of in situ and remote sensing data in near real-time. Within these timings, remote sensing data incoming from the selected instruments are available in their repositories as FITS files and with an L0.5 processing level. This means that these files have not been corrected for instruments response and are in units of digital numbers. In order to calibrate them into physical units it is necessary to achieve the L1 processing level. As a consequence, in order to analyse remote sensing data, this is the first step of the related processing pipeline. This calibration involves correction for the flat field response of the detector, radiometric sensitivity, stray light, geometric distortion, and vignetting. The resulting calibrated images are then archived and analysed for space weather forecast purposes. On the other hand, for our purposes, in situ data do not need calibration. Consequently, this additional step is not necessary.

The calibrated remote sensing data are then analysed by a suitable algorithm and, if the evidence of an event of interest is detected, an algorithm computes the kinematical parameters, the propagation time of the disturbance from the

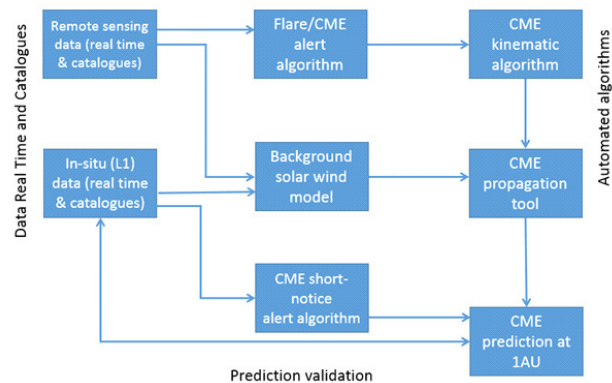


Figure 2: Space weather short-term and medium-term forecast general pipeline at IAU

Sun to the Earth, and the impact probability with the Earth. For instance, if a solar eruption is detected, a CME alert is provided and then the kinematical properties of the observed CME (e.g., initial unprojected velocity and propagation direction) are estimated. Since the CME is observable from different acquired data, these estimated properties are constantly updated, until the observed CME comes out from the last instrument field of view.

At the same time, data incoming from in-situ instruments, which are on-board spacecraft orbiting around the Lagrangian point L1, are constantly retrieved from the dedicated repositories. These ones, together with the remote sensing data related to the heliosphere, are necessary in order to develop an up to date solar wind model (taking into account magneto-frictional forces accelerating/decelerating the eruption and possible CME deflections), which is necessary in order to propagate the eruption up to 1 AU.

As a result, the estimated CME kinematic properties and the solar wind model represents the input for the algorithm that predicts the CME arrival time at L1 and the impact parameter on the Earth.

Finally, the in situ data acquired at L1 are also used to validate the prediction estimated with the remote sensing data.

3. DATA MANAGEMENT

3.1. Data Flow

The first objective of the data management pipeline is near real-time data retrieval from several repositories minimizing latency from source to destination. The data management pipeline is based on crawler components, which query remote repositories for new data products and trigger the internal ingestion and processing pipelines.

The crawling function is implemented to handle many data repositories, some of them providing the same mission data, and to recognize the availability of new data products very soon.

In order to provide services based on the forecast pipeline described in Figure 2, in the current phase of the project the remote sensing and in situ datasets available at the centre and acquired in near-real-time are:

- LASCO C2\C3
- STEREO COR1\COR2
- EUV Imagers: EIT\SWAP
- IN SITU L1 DATA (SOHO, ACE, WIND, DSCOVR)
- OTHER (e.g., geomagnetic, ionospheric data)

In addition, the heliophysics data centre manages several long term archives that are extremely relevant to support the scientific pipeline development and test. The SOLAR mirror archive is an active repository, maintained regularly, while the STEREO archive will be built-in in next months.

The most relevant data for short-term forecast are in situ L1 data (such as local solar wind speed, plasma density, plasma temperature, interplanetary magnetic field, etc...), while for medium-term forecast are remote sensing data (such as EUV full disk images, visible light coronagraphic images, heliospheric images, X-ray fluxes, etc...).

For the medium-term forecast of solar eruption arrival on the Earth, the use of coronagraphic data is mandatory. Nevertheless, there are only a very few space-based coronagraphs providing «near real-time» images of the solar corona nowadays: only two space-based instruments i.e., LASCO and SECCHI.

The Large Angle and Spectrometric CORonagraph (LASCO) is one of the payloads included on the Solar and Heliospheric Observatory (SOHO) spacecraft and it comprises three telescopes (C1, C2, and C3). Only two of them, C2 and C3, are still working. C2 acquires one image (FITS file) every 12 minutes for a total size of 492 MB/day while C3 acquires one image every 24 minutes.

COR1 and COR2 telescopes are part of the Sun Earth Connection Coronal and Heliospheric Investigation (SECCHI) payload on board of one of the two spacecraft of the Solar Terrestrial Relations Observatory (STEREO) mission; i.e., STEREO-A. The instrument takes one triples of images every 5 minutes for a total size of 462 MB/day; every image is related to a code that gives information about the type.

EUV imagers on board of other satellites provide images of the solar disk. For instance, the Atmospheric Imaging Assembly (AIA), payload of the Solar Dynamic Observatory (SDO) acquires images every 12 minutes.

The short-term forecast is based on in situ data. Nowadays three different spacecraft acquire data from the Lagrangian point L1: ACE, WIND and DSCOVR. Unlike data incoming from coronagraphs, that are related to telescopes with different technical features (i.e. field of view) and/or located in different positions, different instruments measure the same physical quantity. For instance, information about the value of three components of the interplanetary magnetic field vector can be retrieved from both, the ACE and the DSCOVR magnetometer.

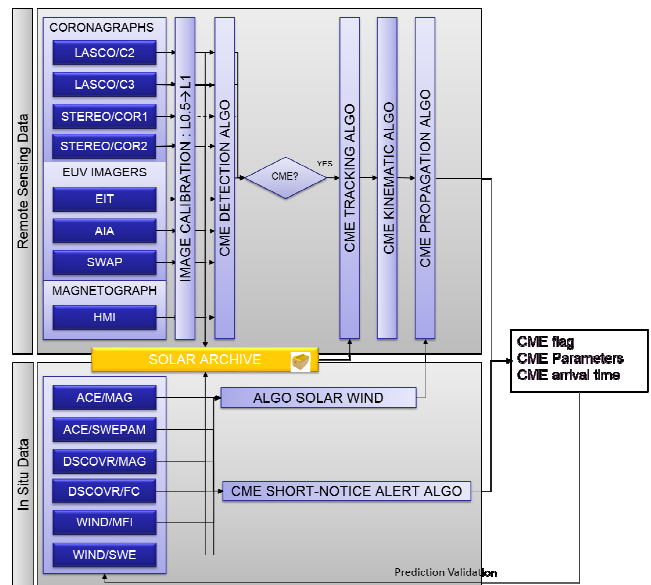


Figure 3: Space weather forecast pipeline at 1AU

For this reason, a cross-check and an integration between these data is necessary. Moreover, the two types of managed data -in situ and remote sensing- presents two other differences: the amount and the size. As a matter of fact, in situ are provided with a higher frequency (about one measurement per minute) with respect to remote sensing data. On the other hand, remote sensing data have a larger total size.

3.2. Data Integration

The integration of many data coming from different instruments is the other important challenge of data management and it is basically obtained through a flexible and scalable design and implementation of a set of interoperable data stores. The first data integration key element is a cross mission data model plus an efficient metadata organization. Metadata are essential to describe, access, and search solar datasets coming from a variety of archives and stored in the original format. There are several data model defined to describe solar and heliophysics data [2][3][4] like VSO, SPASE and ESPAS, among them SPASE is most recently updated that permit to model in a good way data product, instead ESPAS model better the process and the observation from scientific point of view. The proposed data model is based on the SPASE one [3] plus several extension to improve modelling of generated data products and provided services. The usage of space physics ontology to tag data is an improvement already studied in other research projects [4] allowing content-target data search and discovery and machine learning data analysis.

The second issue concerning data integration is the data model mapping and conversion in case of algorithms running on a set of data from different missions. Moreover,

in case algorithms need to process images referred to the same observational period, the system support the time normalization pre-processing task. The latter task together with other pre-processing ones compose a library of pipeline building blocks. The objective is to have an environment that provide not only data but also all tools needed to integrate and use them in order to reduce the time to operations. In such environment scientists can quickly improve their algorithms and they could test them using cross mission dataset, if applicable.

3.3. Architecture

The architecture of the data management system is complex and consists of data stores to archive input data, a metadata data store configured in high availability and processing data stores to prepare data. The input archives are implemented through filesystem based repositories but several technical functional and performance tests will be performed with object storage technologies as these input data is never updated after ingestion. Object storage data store are engineered storage system running on low cost hardware.

For the metadata repository, a technical solution based on high performance search engine such as Elasticsearch is preferred for its level of flexibility. Finally, the implementation of the processing data stores is strictly dependent on applications and kind of data. In-situ scientific pipeline processes time series data thus time series databases are suitable in case of algorithms running on long time range; a possible choice could be Apache Cassandra, which is already used in other projects at ALTEC. Whereas remote sensing data processing is basically image processing, therefore the reference data stores are filesystem or optimized database for images in case of lack of performances.

In order to have an access as much as possible transparent to all data stores the system provides a product manager component managing all different data stores and offering data access services.

Moving to the components dedicated to the processing tasks, they are designed to easily integrate algorithms coded in various languages, and the pipeline can thus be assembled reusing existing and tested routines (e.g., Solar Software is supported). The processing system provides services to track generated data products and algorithms used in processing task in order to support troubleshooting and reprocessing capability. Such complex processing environment is implemented using state-of-art virtualization technologies allowing isolating execution environments and maintaining them under configuration control.

In addition, the heliophysics centre architecture will provide software and hardware components needed to support space weather researches based on a multi-disciplinary approach in order to develop forecast services through machine learning and deep learning techniques.

Although these techniques are not yet used in space weather systems for operational services there is an increasing number of research studies having the objectives to investigate the possibility to recognize unknown patterns or predict physics events mining both historical and new data [5][6][7][8][9].

4. FUTURE PROSPECTS

The future prospects concern algorithms, datasets and services. The data centre will provide the capabilities to evolve current forecast algorithms and develop new ones. In order to reach a finest prediction of the impact time on Earth of the observed phenomenon, non-conventional techniques will be used and implemented in the pipeline.

Moreover, to support new algorithms and improve validation, additional data archives of both existing missions (e.g. SDO) and new missions that will be available in the next years (i.e. Solar Orbiter, Parker Solar Probe, and PROBA-3) will be managed at the Heliospheric Data Centre. In particular, data coming from new space mission designed to set up space weather services (i.e., future L5 ESA mission and Aditya mission in L1) could allow filling up gaps about near real time data availability.

Another possible extension of the data set is represented by data provided from ground-based solar telescope.

Finally, when the project will be in operations, the goals will be to provide medium-term and short-term forecast, data access, reporting and alerting services to Space Weather National Authorities and to integrate with other space weather data centres to compose additional services.

5. REFERENCES

- [1] C.J. Shrijver et al., "Understanding space weather to shield society: A global road map for 2015-2025 commissioned by COSPAR and ILWS", Elsevier Ltd., 2015
- [2] Virtual Solar Observatory Data Model <http://vso.stanford.edu/datamodel/>
- [3] Space Physics Archive Search and Extract (SPASE) <http://www.spase-group.org/>
- [4] ESPAS <https://www.espas-fp7.eu/>
- [5] G. Barnes et al., "A comparison of flare forecasting methods, I: results from the "all-clear" workshop", *The Astrophysical Journal*, 829, 2, 2016
- [6] S.A. Murray et al., "Flare forecasting at the Met Office Space Weather Operations Centre", *Space Weather AGU Journal*, 2017
- [7] M.J. Owens et al., "Probabilistic Solar Wind and Geomagnetic Forecasting Using an Analogue Ensemble or "Similar Day" Approach", *Solar Physics*, Springer, 2017
- [8] O. Olmedo et al., "Automatic Detection and Tracking of Coronal Mass Ejections in Coronagraph Time Series", *Solar Physics*, Springer, 485-499, 2008
- [9] J. Hutton, H. Morgan, "Automated detection of coronal mass ejections in three-dimensions using multi-viewpoint observations", *A&A* 599, A68, 2017

HOW TO APPROACH TO EARTHQUAKE STUDY BY AN INTEGRATED SATELLITE AND GROUND DATA ANALYSIS SYSTEM: THE SAFE ESA-FUNDED PROJECT

A. De Santis¹, C. Abbattista², L. Alfonsi¹, L. Amoroso², M. Carbone², C. Cesaroni¹, G. Cianchini¹, G. De Franceschi¹, Anna De Santis¹, R. Di Giovambattista¹, D. Drimaco², A. Ippolito¹, D. Marchetti¹, F.J. Pavon Carrasco^{1,3}, L. Perrone¹, A. Piscini¹, L. Spogli^{1,4} and F. Santoro²

(1) Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, Rome 00143, Italy
Fax:00390651860308; e-mail: angelo.desantis@ingv.it

(2) Planetek Italia srl, via Massaua 12, Bari

(3) Now at Facultad Física (UCM), Avd. Complutense, s/n. 28040 – Madrid (Spain).

(4) SpacEarth Technology, Via di Vigna Murata 605, 00143, Rome, Italy

ABSTRACT

The SAFE (Swarm for Earthquake study) project aimed at applying the new approach of geosystemics to the analysis of Swarm satellite electromagnetic data for investigating the preparatory phase of earthquakes. This aim requires the integration of a great variety of observations together with satellite data, so the analysis is performed over a big data set. The project has also developed the so-called SAFE Web Platform with the aim of sharing its outcomes and results, and of demonstrating the implemented techniques and algorithms. This contribution presents the SAFE Web Platform with its architecture and functionalities.

Index Terms— Earthquake, Swarm satellites, seismic precursors, Dataset integration, Web-platform

1. INTRODUCTION

The primary goal of the Swarm three satellite mission by ESA is to measure the magnetic signals from the Earth. The SAFE (Swarm for Earthquake study) project (funded by ESA in the framework “STSE Swarm+Innovation”, 2014-2016) aimed at applying the new approach of geosystemics (Figure 1) to the analysis of Swarm satellite electromagnetic data for investigating the preparatory phase of earthquakes [8][1][9].

Swarm constellation consists of three twin satellites that orbit at two different altitudes: two satellites (Alpha and Charlie) fly at lower orbit (around 460 km) while the third satellite (Bravo) flies at around 520 km. Alpha and Charlie fly almost in parallel to improve the knowledge of the East-West gradient of the field.

2. SAFE OBJECTIVE

The main objective of the Project was to explore the possible link between magnetic/ionospheric anomalies and large earthquakes analysing Swarm as well as ground based data (seismic, magnetic, GNSS, etc.). To reach this

objective the SAFE Project dealt with the integrated analysis of more physical parameters whose abnormal variations have been found to be possibly associated with impending earthquakes [2][3][4][5]. These observations are mainly: electromagnetic variations; total electron content and the electron density in the ionosphere, measured both from Swarm satellites and ground-based observatories [6].

The great variety of data analysed (especially those coming from Swarm satellites) is that of a typical big data set, with all consequent implications of this kind of data acquisition and analysis. The particular configuration of the Swarm satellites is expected to provide the best way to possibly detect any electromagnetic anomalous signal produced by strong earthquakes.

3. SAFE WEB PLATFORM

In the framework of the project a dedicated platform has been developed, in order to share information and results with the international scientific community, collect feedbacks on algorithms, and highlight all the project-related events/ initiatives. In addition, the SAFE Web Platform (*safe-swarm.ingv.it*) provides a web application able to automatically collect and update data from catalogue sources, make them available to the user for visualization, and open the possibility to perform customized analyses.

The platform is specially designed to integrate both Swarm-satellites dataset with ground-based geophysical data; it allows for accessing and visualizing Swarm data and performing online user-adjustable analyses.

Considering that the SAFE platform has been designed with the aim to be a powerful analysis and dissemination tool among scientists involved in geophysical studies, the development of a web-based application in this frame offers many advantages over the desktop-based one. For example, when a software application is developed for the web, it could be directly accessed and executed by any user around the world, independently of his running OS (cross-platform

character) and without installing any special software. Only an internet connection and a web browser for executing the

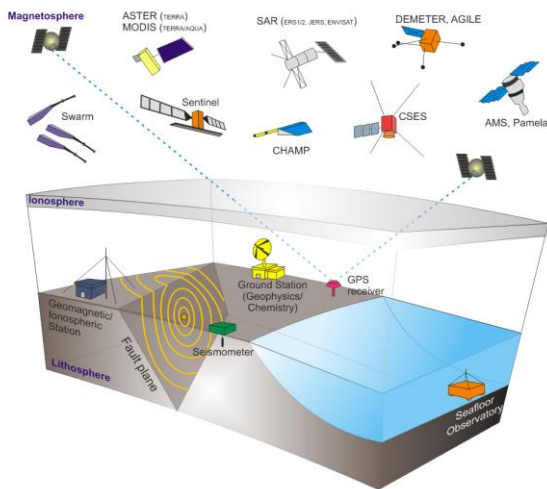


Figure 1. Observations from different platforms (ground, seafloor and satellites) (from De Santis et al., 2015a).

application are needed. This is the main advantage of the web-based applications over the desktop-based ones. In addition, Web applications are centralized in the cloud, in a computer server, so it is guaranteed that the users will always access the newest version of the software, which is a very important feature. More details can be found in the website (safe-swarm.ingv.it).

3.1. Design of the Web Application

The web application designed in this project makes use of different datasets, including data from seismic catalogue or satellite-based data to investigate the existence of electromagnetic precursors in the preparatory phase of an EQ. Thus, the integrated platform has been conceived as a central repository collecting all the input dataset and the results produced by the analysis algorithms.

In particular, the space-born data available in the SAFE platform include both *Level-1B Magnetic* and *Level-1B/2 Ionospheric* datasets (i.e. Langmuir Probe, Plasma dataset, Ionospheric Bubble Index) from Swarm Mission catalogue. At the same time, the ground-based dataset related to EQ events includes data retrieved from the main seismic catalogues (United States Geological Survey-USGS, European Mediterranean Seismological Centre-EMSC, and other national catalogues such as those of INGV and National Observatory of Athens-NOA)

In addition, the platform provides a geographical catalogue, able to display seismic events on the world map, together with the satellite datasets.

Fig.2 shows a simplified scheme representing the design of the system. Thanks to the dedicated workflow engine, the SAFE web-platform is able to lead the complex processing chain involved in the algorithm demonstration, from data

collection, ingestion and catalogue update, until to their processing and results display or delivery to users. Thus, users are able to search, get, view and analyze data from

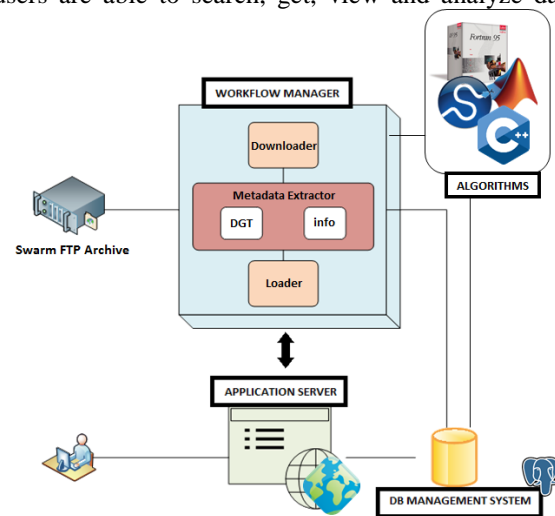


Figure 2. SAFE Web Portal design

different catalogues in a flexible environment. Then they can select the EQ, the kind of analysis, the specific algorithm and perform the desired analyses.

This is achieved by means of programmable components organized in different layers, according to a simplified Service-Oriented Architecture (SOA) model.

Thanks to its design, the integration of new functionalities in the system is simplified. In particular, the SOA is implemented with the well assessed 3-Tier architectural design pattern, so including **Data**, **Business Logic** and **Presentation** layers.

The catalogue functionality allows navigating, filtering and visualizing Swarm data-ground tracks per product (for all of three satellites, Alpha, Bravo, and Charlie). An automatic harvesting mechanism has been implemented exploiting the FTP-based interface provided by ESA for data download. This harvesting mechanism potentially provides the capability of (automatically) keeping daily aligned the SAFE Swarm products database to their actual availability. In this sense, this platform practically offers a graphical catalogue of Swarm mission products.

Furthermore, directly from within the geographical viewer, once an earthquake event is selected, users are able to perform their own analysis choosing an algorithm and sending a request through the web browser. Due to algorithms' novelty and complexity, they have been integrated in their original prototypal form and in a high-level scientific data processing environment (Matlab®, Scientific Python, or FORTRAN), while few components are engineered in C++.

Main technologies exploited in the SAFE platform are shown in Fig. 3: the backend side makes use of dedicated tools, such as Flask (Python micro-framework for web-

applications [10]) and Redis (key-value storage database for queues hosting [11][12]), while the frontend side is

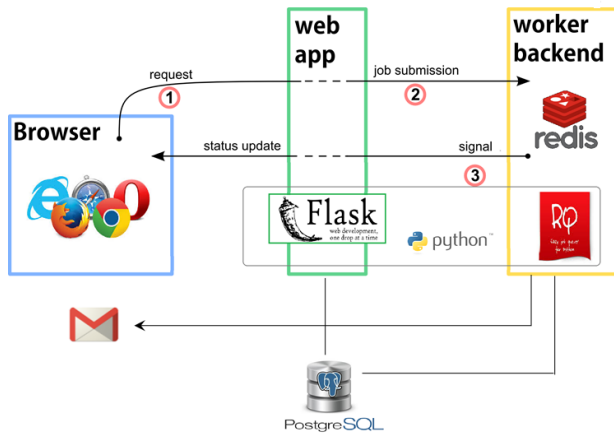


Figure 3. Main technologies and data flow in SAFE Web Platform.

implemented using Javascript and AJAX technologies. In addition, the remote server uses PHP programming language to perform the requested operations, and a Postgres database is used to store the data.

The Web application layer submits user requests to the remote server, decoupling client interface from backend worker that in turn interacts with the database; finally, backend worker responds to client layer queuing job and updating request status.

Then, the choice of making available all the functionalities as web services allowed for the integration with the project website, providing also the possibility to practically demonstrate algorithms results on the selected test cases.

3.2. User Interface and Functionalities

User interface (UI) is a key element as it represents the communication point between the user and the system. In general, our aim was to develop a Web application that allows the user to easily navigate both satellite-based and ground-based catalogues, and perform customizable analysis on a selected dataset.

The SAFE Web Portal homepage allows to access to every system section, according to different permissions assigned both to science team members and to generic users by means of an appropriate authentication mechanism and of the data integrity and availability.

The main sections of the Web Portal are:

- **Project:** it contains general information about the project status;
- **News & Events:** it contains last news and event related to the project;
- **Resources:** it contains main scientific results related to the SAFE project, and links to SAFE Viewer and Catalogue;

- **Audience:** it contains information about methods and analyses used in the frame of SAFE-project, for both Scientist user and Private/Public stakeholders;

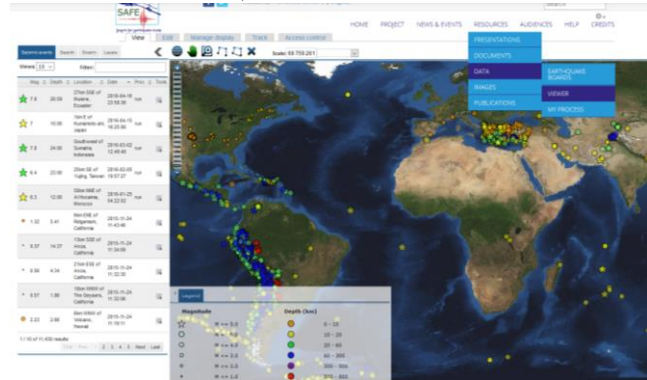


Figure 4. SAFE Viewer Interface; seismic events available in the seismic catalogue are listed on the left side.

General sections of the website containing details about the project and program of events are public and visible to everyone. In addition, the system includes a registration functionality and authentication with different permission levels in order to access to reserved sections.

The system allows managing and visualizing data and products related to the available datasets, through a search interface available in the Viewer section from the Resources Menu, as shown in Figure 4.

Earthquake events available in the seismic dataset are listed in the Viewer section (see Fig. 4, left) and catalogued according to the most important metadata (i.e. date, magnitude, hypocenter depth, epicenter geographical coordinate, etc.). In addition, each event is visible on the map, achieved by Geoserver [13] technology, according to labels (*Magnitude, Epicenter Depth*) reported in the Legend window. Users are able to navigate the EQ list and search for a specific event via an interactive tool, giving the possibility to integrate maps of seismic events and Swarm satellites ground-tracks in a unique view, allowing for the synoptic visualization of all available datasets.

When the user selects an event in the Viewer section, the system shows all the associated metadata in the summary window at the bottom of the page (see Figure 5). Using the same window, scientists are able to launch a process to study the selected event and run some specific algorithm. In particular, the user can choose one of the available analysis algorithms developed by the research team involved in the SAFE Project.

Each algorithm is associated to a set of input parameters that can be configured using the web interface; after the selection of input parameters, the process can be started clicking the *Run* button (see Fig. 5).

Depending on task duration, the system performs each analysis as synchronous or asynchronous process. For

synchronous jobs, the user is able to view the generated output in Web Application and download it immediately.

On the contrary, when “long” tasks end, the system sends to the user an email containing the link for the download of job results.

An additional page named **MyProcess** (reserved to each single user), is available in the **Resources** section. This page provides details about the processes launched by each user (i.e. Algorithm Name, Study-field, StartDate, EndDate, etc.) and allows jobs monitoring. In addition, users are able to download analysis results or visualize a results preview.

This approach allows the users to have a synoptic vision of potentially related datasets, and, moreover, to launch a user-customized analysis using events available in the catalogue.

4. CONCLUSIONS

The usefulness of the presented Web platform has been demonstrated in several case studies (please connect to safeswarm.ingv.it to see the different cases) where multiple analyses of geomagnetic, ionospheric and seismic data are performed in order to investigate the process of earthquake preparation in the frame of Lithosphere-Atmosphere-Ionosphere Coupling (LAIC) theory [7].

Furthermore, the presented technologies and architectural design make the SAFE Web Platform a very flexible application able to implement and test the functionalities identified as users’ needs in the SAFE project. It also allows for resources scalability going towards real-time scientific analyses and supporting operators in large data processing tasks. In this way, the capability to view and analyze different data in an integrated environment provided with social and web capabilities will certainly favor the growth and the spread of the academic debate in the field of geo-seismological research.

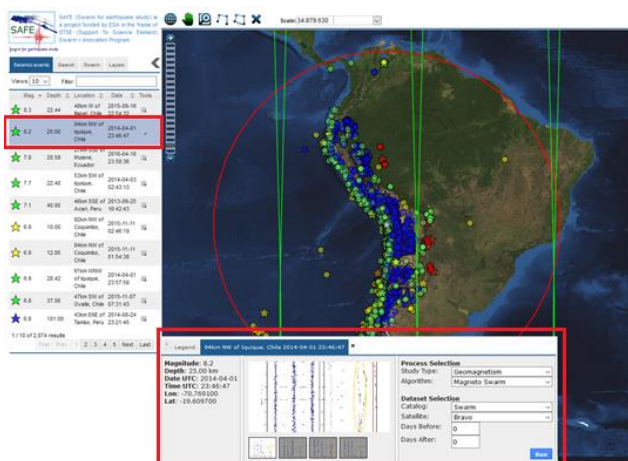


Figure 5. SAFE Viewer, On-line Analysis Tool Interface. Additional features are shown on the map according to the selected event: the Dobrovolsky Area (red circle) and satellite data ground-track (green lines) related to EQ day.

5. ACKNOWLEDGEMENTS

The SAFE web portal has been developed by Planetek Italia in the frame of the SAFE Project, funded by the European Space Agency (Contract No.4000116832/15/NL/MP) under the STSE (Support To Science Element) Swarm+Innovation Program, and coordinated by the INGV - Istituto Nazionale di Geofisica e Vulcanologia.

6. REFERENCES

- [1] Cicerone, R.D., Ebel, J.E., Britton, J., A Systematic Compilation of Earthquake Precursors. *Tectonophysics* 476 (3-4). Elsevier B.V.: 371–96. doi:10.1016/j.tecto.2009.06.008, 2009.
- [2] De Santis A., Cianchini G., Qamili E., Frepoli A.. The 2009 L'Aquila (Central Italy) seismic sequence as a chaotic process, *Tectonophysics*, 496 44–52, 2010.
- [3] De Santis A., Cianchini G., Beranzoli L., Favali P., Boschi E., The Gutenberg-Richter law and Entropy of earthquakes: two case studies in Central Italy, *BSSA*, v.101, 1386-1395, 2011.
- [4] De Santis A. et al., Geospace perturbations induced by the Earth: the state of the art and future trends, *Phys. & Chem. Earth*, 85-86, 17-33, 2015a.
- [5] De Santis, A., Cianchini, G., Di Giovambattista, R., Accelerating moment release revisited: Examples of application to Italian seismic sequences, *Tectonophysics*, Vol. 639, 82-98, 2015b.
- [6] De Santis A. et al., Potential earthquake precursory pattern from space: the 2015 Nepal event as seen by magnetic Swarm satellites, *Earth and Planetary Science Letters*, 461, 119-126, 2017.
- [7] Pulinet S, Ouzounov, D., Lithosphere-Atmosphere-Ionosphere coupling (LAIC) model: an unified concept for earthquake precursors validation. *J. Asian Earth Sci*, 41(4–5):371–382, 2011.
- [8] Scholz, C.H., *The Mechanics of Earthquake and Faulting*, vol. xxiv. Cambridge Univ. Press, Cambridge/New York, 471 pp, 2002.
- [9] Wyss M. (ed.) *Evaluation of proposed earthquake precursors*, American Geophys. Union, 1991.
- [10] Flask: web microframework for Python, Documentation, <<http://flask.pocoo.org/docs/0.12/>> [December 2016]
- [11] Programming with Redis, Documentation, <<https://redis.io/documentation>> [December 2016]
- [12] RQ: easy job queues for Python, Documentation, <<http://python-rq.org/docs/>> [December 2016]
- [13] Geoserver User Manual, <<http://docs.geoserver.org/>>

HARNESSING BIG DATA FOR AGRICULTURE MONITORING : COMBINING REMOTE SENSING, OPEN ACCESS DATA AND CROWDSOURCING

Raphaël d'Andrimont^{1,*}, Guido Lemoine¹, Marijn van der Velde¹, Christina Corbane²

European Commission, Joint Research Centre (JRC),

¹ Sustainable Resources Directorate, Food Security Unit (D.5),

² Space, Security and Migration Directorate, Disaster Risk Management (E.1)

Via E. Fermi 2749, 21027 Ispra, Italy

*raphael.dandrimont@ec.europa.eu

ABSTRACT

Timely agriculture monitoring at parcel level is becoming a reality at continental scale. The recent deployment of the Sentinel fleet has paved the way for new developments ranging from the creation of farm services to the redrawing of agricultural policies. However, suitable processing capacities, robust algorithm development along with systematic ground data collection are still required to build an operational monitoring system at parcel level. By reviewing the user needs and the recent achievements, this paper questions how agriculture monitoring could benefit from a synergistic use of evolving big data solutions from Earth Observation processing, open access reference data, and latent information derived from crowdsourcing.

Index Terms— Crop, citizen science, Sentinel

1. INTRODUCTION

Earth Observation (EO) data are growing in size and variety at an exceptionally fast rate. New satellites, airborne, and ground-based remote sensing systems characterized by high spatial, temporal and radiometric resolution are available. The term Big Data is one of the current major trends in data science and Information Communication Technology (ICT). However the Big Data concept itself is elusive, bringing too many possible definitions according to specific applications that are foreseen [1]. A now accepted consensus definition of Big Data proposed by [2] is characterized by the 5 "V"s: *volume*, *variety*, *velocity*, *veracity* and *value*. Using this definition, we evaluate how the *volume*, *variety* and *velocity* are currently being addressed while the *value* and *veracity* are still missing for most agricultural Big Data applications.

2. USER NEEDS IN AGRICULTURE

EO data use may address a hierarchy of different user categories in the agricultural sector, as abstracted in Figure 1. At

the basis are individual farmers, who can use EO inputs in farm management applications (e.g. precision farming) and regulatory contexts (e.g. aid applications, insurance claims). Farm service industry and regional authorities require information at administrative or thematic aggregation levels that is relevant for their line of business (e.g. a regional supplier of farm inputs, a machinery operator and a local watershed management department). At regional or national level, food processing industry and administrations require further aggregated statistics, e.g. for storage and transport logistics, national policy monitoring. At the top of the pyramid, organizations active in international trade, food security assessments, and agricultural policy development are key users of (aggregated) crop production statistics. Currently, EO use is most established amongst the higher levels of the agricultural user pyramid. Obviously, levels of knowledge detail depend, across sectors, on other technology factors such as ICT technology uptake, complementarity with existing information and reliability of the delivery of the information at the right level of specificity for each sector.

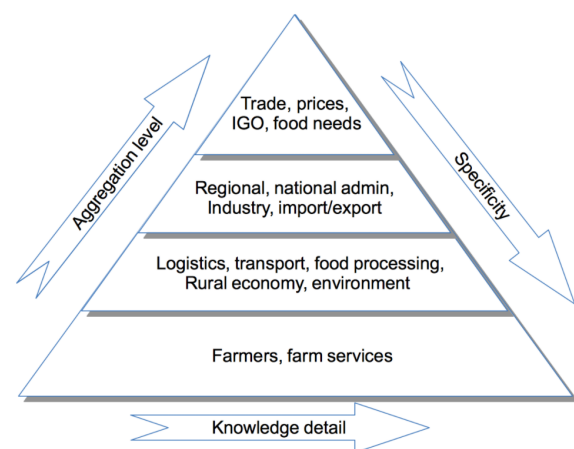


Fig. 1. EO use across the agricultural user community.

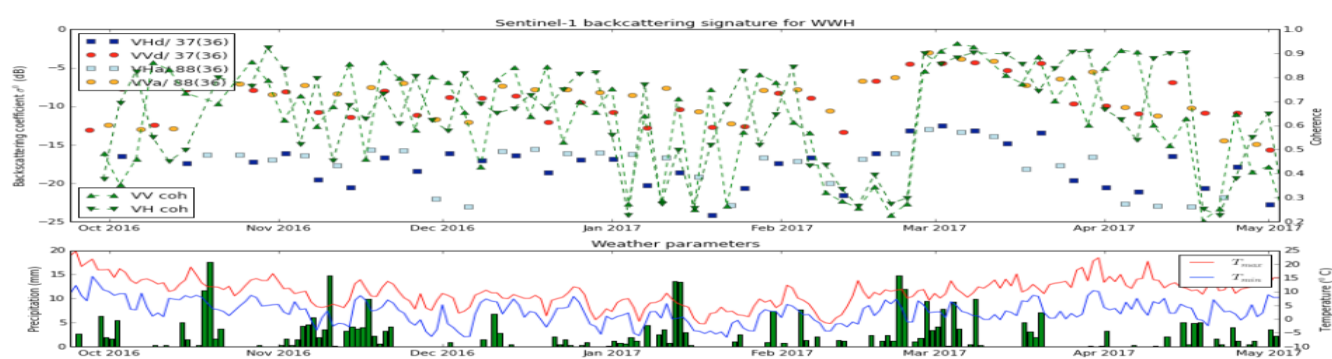


Fig. 2. Multi-temporal Sentinel-1 profiles for a winter wheat parcel in Noordoostpolder, the Netherlands. The seasonal trend in backscattering coefficients and coherence relate to soil cultivation, crop growth and harvesting. Weather data is needed in support to noise filtering and data interpretation. We use Sentinel-1 signatures extensively in crop mapping.

Essential developments in open access to reference data are instrumental to enable rapid integration of EO derived information. In Europe, the reference data sets required for the management and monitoring of the Common Agricultural Policy (CAP) are now often released in the public domain. At the same time, agricultural production systems are increasingly relying on ICT solutions (sensors networks, GPS, electronic tracking and tracing, guided machine operations, etc.). Facilitated by mobile communication, there is unprecedented potential to accrue mutual benefits through intensified information sharing across the food production and processing industry.

3. DATA PARADIGM SHIFT: VOLUME, VARIETY AND VELOCITY

At the same time, the Copernicus Data Policy has provided open and free access to Sentinel products for any user. In constellation, Sentinel-1, -2 and -3 have a revisit capacity of respectively 6 days, 5 days and 1-2 days with a finest spatial resolution of respectively 10 m, 10 m and 300 m. The volume of Sentinel data available to users will increase at a rate of approximately 10 PB per year. In Figure 2 we demonstrate the depth of the multi-temporal data stack for Sentinel-1. Sentinel-1 acquisition density over the entire European continent has reached an unprecedented level since the start of operations of Sentinel-1B in October 2016. Combination of ascending and descending and overlapping orbits provide a consistent data stream with a 2-3 days rhythm for most European locations. Furthermore, backscattering coefficients (in VV, VH) can now be combined with coherence, generated from same-orbit S1A and S1B pairs, every 6 days. Figure 2 also provides a good impression on the challenges we face in visualization and interpretation. We use these data already extensively for automated crop mapping and monitoring.

To be able to use this data deluge, processing capacities are being developed addressing the *volume*, *variety* and *velocity* challenge of the Big Data. Several platforms are or will

be available in the future to process the massive Copernicus Sentinel data stream.

Started in 2010, Google Earth Engine (GEE) was the first cloud-based platform exploiting Sentinel data for planetary-scale geospatial analysis using massive computational power. GEE demonstrated its ability to tackle a variety of high-impact societal issues including deforestation, drought, disaster, disease, food security, water management, climate monitoring and environmental protection [3].

Recently, the European Commission has envisaged the creation of the Copernicus Data and Information Access Services Operations (DIAS). By providing data and information access alongside processing resources, this initiative is expected to boost user uptake, stimulate innovation and the creation of new business models based on EO data and information. The DIAS should deploy operational access platforms in early 2018. Started in 2016, the Joint Research Centre Earth Observation Data and Processing Platform (JEODPP) [4] has a large-scale storage coupled to high computing capacities. The JEODPP access is currently limited to the Joint Research Centre but could be used as a proof of concept for DIAS development. The automatic download of Sentinel data is triggered for pre-defined areas of interest and the infrastructure has been successfully tested in several user domains, such as crop classification, maritime surveillance and forest mapping.

4. USING LATENT INFORMATION TO BRING VALUE AND VERACITY

Now that the prerequisite features including *volume*, *variety* and *velocity* are being addressed, it is necessary to develop operational applications to integrate the *value* and *veracity* components into the data stream. Two different information sources are currently being tested.

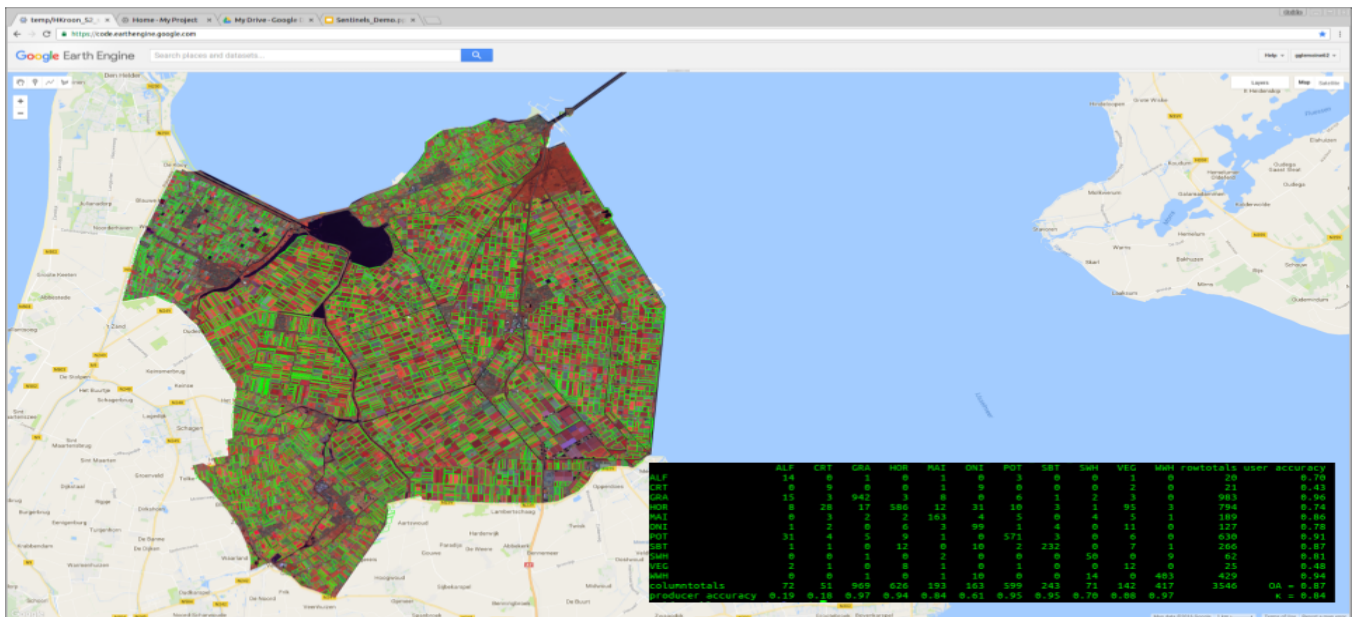


Fig. 3. Example of the use of full cover Dutch LPIS-based parcel declaration in a Random Forest classification experiment using Sentinel-2A in Google Earth Engine.

4.1. Open-access Land Parcel Identification System (LPIS)

In the European Union, the Member States' Land Parcel Identification Systems (LPIS), i.e. very high spatial agricultural parcel maps to aid the management of the Common Agricultural Policy (CAP), are being released as open access data in an increasing number of regions. LPIS contain up to several millions parcel boundary vectors, depending on Member State, delineating agricultural land eligible for CAP support. Farmers use the LPIS to declare their cropping practices, including specific environmental measures where relevant, in an annual aid application. Member State administrations maintain LPIS and carry out management and control activities on the basis of the LPIS, often with the use of remote sensing¹.

The level of detail available in the various open access LPIS implementations varies across Member States. For instance, in the Netherlands² and Austria³, the actual parcel declarations are made public annually, whereas, for instance, Denmark⁴ and the Czech Republic publish only the LPIS reference vectors.

Either of the data sets provides essential information on agricultural landscapes which can be integrated with satellite imagery data sets for crop mapping purposes, crop status monitoring and, in the context of CAP management and control, providing information on compliance with the CAP conditions on eligible aid. Obviously, with these data sets, is-

¹<https://marswiki.jrc.ec.europa.eu/>

²Basis Registratie Gewaspercelen

³ INVEKOS Schlge

⁴Jordbrugs analyser

sues on robustness of classification approaches, transferability of trained machine learning models, cross-sensor performance evaluation, etc. can be tested at scales that matter at the level of full regions or countries, opposed to a small pre-conditioned study area (see Figure 3). For now, we work with these data sets on a country-by-country basis in Europe, due to the considerable variation in data quality and detail, but also look into harmonization efforts to ensure that these data sets as easily accessible for big data analytics for agricultural use cases.

4.2. Collecting *in situ* data

For the verification and validation of robust algorithms with deep Sentinel data stacks and very large parcel reference data, high quality and actual *in situ* data collection is required. Exploratory studies using citizen observatories or crowdsourcing already demonstrated their potential to collect massive high-quality *in situ* data. Citizens without professional expertise in remote sensing can become actively involved in the creation and analysis of large data sets, which is known as crowdsourcing. The rise of geospatial user-created content such as Geo-Wiki [5] has been of great benefit to the collection of large quantities of reference data for land cover classification, involving citizens in the collection of validation data. We are exploring their potential in synergy with remote sensing to obtain information about crop phenology and agricultural practices (Figure 4). More specifically, we are exploring the potential of active and opportunistic crowdsourcing.

In our case, the active crowdsourcing consists of setting

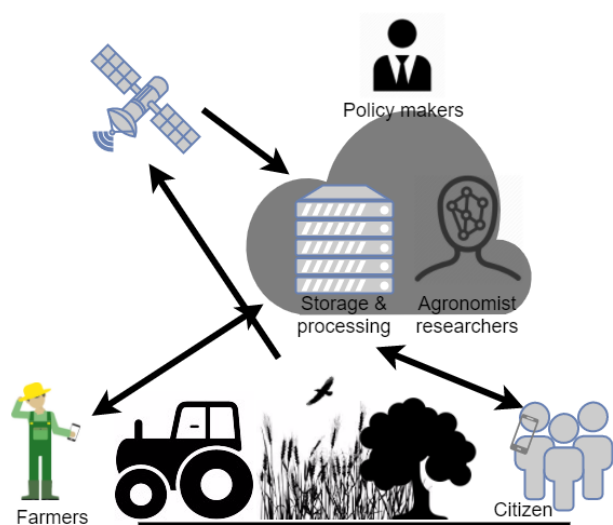


Fig. 4. Monitoring crop at parcel level by adding *value* and *veracity* to the data thanks to in-situ crowdsourcing collection.

up field sampling based on *a-priori* information combining LPIS data and satellite time series to monitor agricultural practices and phenology events during crop growth such as ploughing, mowing, emergence, flowering and harvest. This information is then used to draw a sample and identify the fields targeted for *in situ* data collection. The data collection could be done with mobile applications where people could be guided to particular fields to take pictures and observations such as in FotoQuest⁵ or with an ad-hoc crop management application⁶. Alternatively, we are testing the collection of geolocated street-level photographs. To this end, we are using and contributing to the Mapillary⁷ platform which is the first platform to provide open access and free-of-charge detailed street photos based on crowdsourcing [6].

By contrast, opportunistic crowdsourcing uses information readily available on social media, such as Twitter, to obtain timely information about crop development. Both approaches could unleash the power of Big Data for agricultural monitoring by bringing them *value* and *veracity*.

5. CONCLUSIONS

The high spatial resolution of the Copernicus Sentinel data (~ 10 m), short revisit cycles (a few days), extraordinary orbit stability, and near contemporaneous cross-sensor observations, combined with appropriate data storage and processing facilities are now a reality. The unprecedented data stream must now be matched with rapid uptake in meaningful applications. The potential of such applications in the agricultural user domain is significant. However, ground-based data

streams are needed to transform our capacity to monitor crop conditions at parcel level and scale these to meaningful production areas. Many challenges remain, such as spatially and coherently aggregating heterogeneous parcel-based information across scales. Yet, developments that foster open access to detailed geospatial reference datasets, and the exploitation of active and passive crowdsourcing, can enable a much better characterization of processes, conditions, and impacts on crops in the field. A key challenge is to involve the various actors across the agricultural user pyramid in rich and systematic information exchange by creating novel incentives that mutually benefit those actors. Consistent information derived from Copernicus Sentinel data has the potential to become the currency in such exchanges. This requires a significant step up from the typical experimental set up in scientific experiments to realistic regional, national or even continental scales. Big Data Analytics is the enabling technology to achieve this step up.

6. REFERENCES

- [1] Stefano Nativi, Paolo Mazzetti, Mattia Santoro, Fabrizio Papeschi, Max Craglia, and Osamu Ochiai, "Big data challenges in building the global earth observation system of systems," *Environmental Modelling & Software*, vol. 68, pp. 1–26, 2015.
- [2] Mark A Beyer and Douglas Laney, "The importance of big data: a definition," *Stamford, CT: Gartner*, pp. 2014–2018, 2012.
- [3] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore, "Google earth engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017.
- [4] P Soille, A Burger, D Rodriguez, V Syrris, and V Vasilev, "Towards a jrc earth observation data and processing platform," in *Proceedings of the Conference on Big Data from Space (BiDS16)*, Santa Cruz de Tenerife, Spain, 2016, pp. 15–17.
- [5] Steffen Fritz, Linda See, Ian McCallum, Liangzhi You, Andriy Bun, Elena Moltchanova, Martina Duerauer, Fransızka Albrecht, Christian Schill, Christoph Perger, et al., "Mapping global cropland and field size," *Global change biology*, vol. 21, no. 5, pp. 1980–1992, 2015.
- [6] Levente Juhász and Hartwig H Hochmair, "User contribution patterns and completeness evaluation of mapillary, a crowdsourced street level photo service," *Transactions in GIS*, vol. 20, no. 6, pp. 925–947, 2016.

⁵<http://fotoquest-go.org/>

⁶Such as <https://landsense.inosens.rs>

⁷www.mapillary.com

NEXT STEP FOR BIG DATA INFRASTRUCTURE AND ANALYTICS FOR THE SURVEILLANCE OF THE MARITIME TRAFFIC FROM AIS & SENTINEL SATELLITE DATA STREAMS

*R. Fablet¹, N. Bellec¹, L. Chapel³, C. Friguet³, R. Garello¹, P. Gloaguen³, G. Hajduch⁴, S. Lefèvre³
F. Merciol³, P. Morillon², C. Morin⁵, M. Simonin⁵, R. Tavenard⁶, C. Tedeschi², R. Vaddaine⁴*

¹ IMT Atlantique, Lab-STICC, UBL ² Univ. de Rennes 1, IRISA-MYRIADS ³ Univ. de Bretagne-Sud, IRISA-OBELIX

⁴ CLS, Espace et Segments Sol, Brest⁵ Inria, IRISA-MYRIADS⁶ Univ. de Rennes 2, IRISA-OBELIX

ABSTRACT

The surveillance of the maritime traffic is a major issue for security and monitoring issues. Spaceborne technologies, especially satellite AIS ship tracking and high-resolution imaging, open new avenues to address these issues. Current operational systems cannot fully benefit from the available and upcoming multi-source data streams. In this context, SESAME initiative aims to develop new big-data-oriented approaches to deliver novel solutions for the management, analysis and visualisation of multi-source satellite data streams going beyond the CLS implementation. Targeted at the automatic generation and documenting of early warnings, our key originality lies in a big-data approach to jointly address these challenges based on the complementarity of the scientific and operational expertise gathered in the consortium: big-data platforms, mining strategies for time series and trajectory data, Sat-AIS signal analysis, high-resolution satellite imaging.

Index Terms— Sentinel, high-resolution satellite imaging, AIS maritime traffic surveillance, big data, data mining, behaviour analysis, ship detection.

1. CONTEXT AND CHALLENGES

The surveillance of the maritime traffic is a major issue for maritime security (e.g., traffic monitoring, border surveillance, ...) as well as environmental monitoring (oil spill monitoring, ...) and marine resource management (illegal fishing monitoring, ...). Spaceborne technologies, especially satellite ship tracking from AIS messages (Automatic Identification System) and high-resolution imaging of sea surface, open new avenues to address such monitoring and surveillance objectives. Currently, operational systems cannot fully process these complete streams of satellite-derived data. For instance, the French institutional users evaluate that less than 20% of the overall AIS data (about a few tens of millions of AIS messages daily) are actually analysed for abnormal behaviour detection. Besides, the free access to Sentinel Earth Observation data streams (high-resolution Sentinel-1 SAR

and Sentinel-2 optical imaging, up to a few TB daily [1]) offers novel opportunities for the analysis and detection of ship behaviours, including AIS-Sentinel data synergies.

In this context, SESAME initiative aims to develop, implement and evaluate new big-data-oriented approaches to deliver novel solutions for the management, analysis and visualisation of multi-source satellite data streams. Targeted at the automatic generation and documenting of early warnings (both in real-time and re-analysis modes), the key scientific and technological challenges cover the development of hardware and software platforms adapted to the characteristics of the data streams of interest along with the design of novel models and algorithms for AIS-Sentinel synergies and the automatic detection of abnormal behaviours. Besides the development of CLS big data infrastructure [2], the originality of the project lies in a big-data approach to jointly address these challenges based on the complementarity of the scientific expertise and knowledge of the operational needs gathered in the consortium: big-data platforms, mining strategies for time series, modeling and analysis of tracking data, Sat-AIS signal analysis, high-resolution satellite imaging.

2. RELATED WORK

Recent works highlighted the importance of new approaches for knowledge discovery and anomaly detection in maritime surveillance [3, 4, 5]. The use of AIS data becomes more and more important to improve traffic monitoring. Practical methods were developed to detect data falsification or spoofing [6] or to detect abnormal behaviours in trajectories [8, 7], both to detect dangerous movement patterns or potentially illegal behaviours. Yet, the combination of such rule-based and statistical models to big-data infrastructure and frameworks largely remains to be investigated. Besides, the emergence of deep learning techniques [9, 10] is particularly appealing to further investigate large-scale AIS data streams.

Ship detection from satellite imagery has long been an active research topic (e.g. [11]). Synergies between AIS and Sentinel data streams appear promising [5, 12] to improve

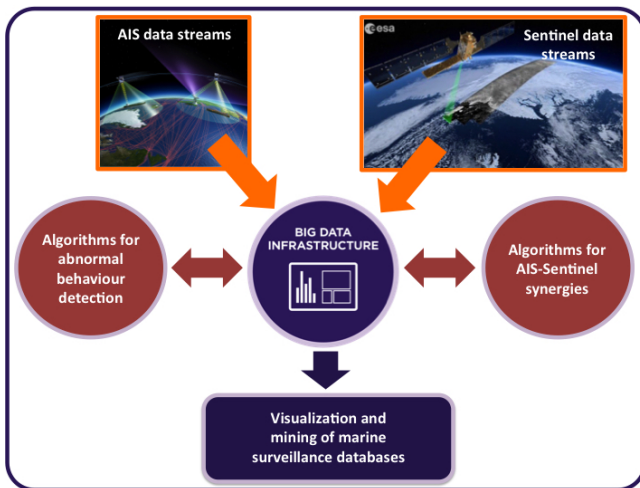


Fig. 1. SESAME workflow for big-data-oriented maritime surveillance from multi-source AIS and Sentinel data streams.

surveillance performance and double check the cooperative AIS data with non-cooperative sensors. We propose to unify these two approaches within a big data framework.

3. PROPOSED WORKFLOW

As sketched in Fig.1, the proposed workflow relies on the implementation and evaluation of a big-data-oriented framework for the management, mining and visualisation of the considered multi-source data streams. It combines the development of hardware and software big data platforms with novel models and algorithms for the detection of abnormal behaviours and AIS-Sentinel synergies. SESAME embeds the implementation of the proposed solutions for dual case-studies for the automatic generation and documenting of early warnings: the real-time analysis of AIS-Sentinel data streams and the re-analysis of large-scale AIS-Sentinel datasets. They will comprise both the evaluation of algorithms and models as well as big-data-oriented infrastructures and frameworks. Grid'5000 platform[14] will provide an initial flexible and scalable testbed, whereas CLS big data infrastructure will be the targeted platform to validate how the proposed solutions scale-up to realistic large-scale datasets.

The algorithms and models will be evaluated on two case studies representative of the challenges of global maritime surveillance. The first case study will focus on the monitoring of IUU (Illegal, Unreported, and Unregulated) fishing activities occurring inside the economic exclusive zones (EEZ). A more general approach will be taken for the second case study where the maritime traffic and illegal activities will be analyzed on a global scale.

4. PRELIMINARY RESULTS

4.1. Big data framework for AIS data streams

CLS operational rule-based architecture for the mining of AIS data streams will provide a baseline architecture for the development and evaluation of big-data-oriented infrastructure. Preliminary analyses point out the requirement for the exploitation of big-data-oriented frameworks to scale up to large-scale AIS datasets. Grid'5000 is used as a pre-production testbed to evaluate the framework.

Fig. 2 depicts the envisioned pipeline of computation and storage for the AIS data. It is split into two parts: first, the real-time stage where data streams are analyzed in near real-time, and second, the batch stage where larger history of data can be processed at once. Outputs of these two stages can either be stored persistently, pushed to other message queues for future reuse or can be alerts passed to a human analyst for further analysis. The different processing components in this architecture rely on a message brokering system taking care of messages buffering and distribution between the various actors. This architecture is extensible in the sense that new components (e.g corresponding to new types of alerts) can be added as well as new storage backends (e.g for indexing purpose).

Our initial investigations focused on the characterization of the AIS data stream. In particular, based on one month of AIS data on a global scale, we started investigating the unsortedness in AIS message arrivals. We compared the order in which AIS messages has been stored in the storage system versus its sending timestamp. This will in turn help us replay the global AIS stream realistically for simulation and validation purposes. The initial platform version was based on a Spark cluster¹ for the batch processing system and Kafka² for the message brokering and real-time processing system. The platform deployment is fully automated and run on the Grid'5000 platform [14].

4.2. Synergy between Spaceborne SAR data and AIS data

Regarding the joint usage of spaceborne SAR data and AIS information, CLS has demonstrated its relevance for the characterisation of SAR vessel detection performances using interpolated/extrapolated AIS tracks as ground truth [11]. It is operationally used for the monitoring of oil spills at sea and the identification of potential polluter sources [5] in the framework of the European Maritime Safety Agency (EMSA) CleanSeaNet Service [13].

Preliminary results highlight the relevance of the synergy between SAR and AIS information in terms of geographic coverage and of detection and characterisation of abnormal activities at sea. We also demonstrated the feasibility of

¹[urlhttps://spark.apache.org/](https://spark.apache.org/)

²<https://kafka.apache.org>

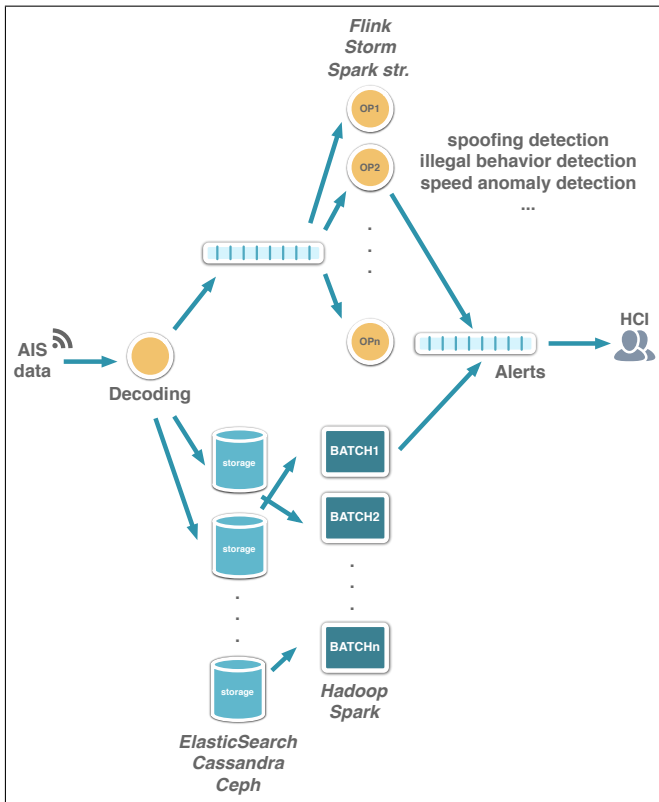


Fig. 2. Targeted SESAME logical infrastructure: it aims to generate alerts for abnormal behaviours from the AIS data stream. As illustrated, the infrastructure will exploit and combine state-of-the-art big-data-oriented framework such as for instance Cassandra, Hadoop, Spark, Fink,...

the construction of large-scale datasets of SAR echoes acquired in various configuration corresponding to a subset of known vessels with a view to applying machine-learning-based detection and classification strategies. Considering a four-month dataset of Sentinel-1 A satellite data over Europe from March to June 2017, we collected 5414 SAR images. They were systematically processed using CLS vessel detection algorithm. The detected SAR echoes were then matched with interpolated/extrapolated AIS data. For illustration purposes, we depict in Fig. 3 the spatial mapping of the cumulated number of SAR-AIS matches over the considered 4-month period. These results highlight shipping routes and areas where the coverage of the AIS network is significantly lower than in densely-monitored areas such as in the English channel. The Bay of Biscay as well the northern coast of Tunisia are typical examples of areas with a high-traffic but a relatively low coverage in terms of coastal AIS receiver networks. In such areas, the main source of AIS data is issued from satellite AIS data, but it cannot provide a space-time coverage similar to dense coastal AIS networks. Our future work will further explore the potential of AIS-SAR synergies

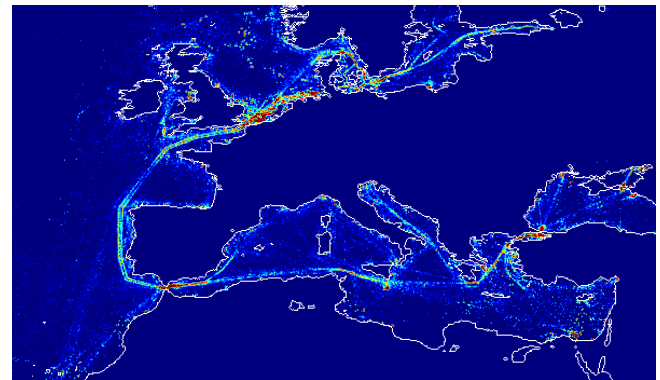


Fig. 3. Maps of the cumulated number of AIS messages matched with SAR echoes from March to June 2017: the redder the color, the higher the density of AIS-SAR matches. For this analysis, we processed 5414 Sentinel-1 A satellite images using CLS operational vessel detection system.

for the creation of groundtruthed SAR image datasets for the development of novel learning-based ship vessel detection and identification strategies.

5. CONCLUSION

The expected impacts of the project include both dissemination actions to the scientific community, including a maritime surveillance benchmark suite, and technological transfers to CLS with respect to future national and international calls on operational systems and services for maritime traffic surveillance and high-resolution environment monitoring.

6. ACKNOWLEDGEMENTS

The authors acknowledge the support of DGA and ANR under reference ANR-16-ASTR-0026 (SESAME initiative), the labex Cominlabs, the GIS BRETTEL (CPER/FEDER framework) the Booster MoreSpace and CNES for access to the PEPS cluster.

7. REFERENCES

- [1] A.G. Castriotta, "Sentinel data access report 2016.," Tech. Rep.
- [2] E. Lambert et al, "Hadoop platform for georeferenced mobiles and gridded data," in *Submitted to Big Data From Space 2017*, 2017.
- [3] FP7 DOLPHIN Consortium, "FP7 DOLPHIN Project," <http://gmes-dolphin.eu/>,
- [4] "Maritime Knowledge Discovery and Anomaly Detection Workshop," Ispra, Italy, July 2016, European Commission, pp. 24–27 – ISBN 978–92–79–61301–2.

- [5] N. Longép   et al, “Polluter identification with space-borne radar imagery, ais and forward drift modeling,” *Marine Pollution Bulletin*, vol. 101, no. Issue 2, pp. 826–833, 2015.
- [6] C. Ray et al, “Methodology for Real-Time Detection of AIS Falsification,” in *Maritime Knowledge Discovery and Anomaly Detection Workshop*, Michele Vespe and Fabio Mazzarella, Eds., Ispra, Italy, July 2016, pp. 74–77.
- [7] G. Pallotta et al, “Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction,” *Entropy*, 15(6):2218–2245, 2013.
- [8] W. Hu, et al, “A system for learning statistical motion patterns,” *IEEE PAMI*, 28(9):1450–1464, 2006.
- [9] Y. Le Cun, et al, “Deep learning,” *Nature*, 521(7553):436-444, 2015.
- [10] X. Jiang et al, “Trajectorynet: An embedded gps trajectory representation for point-based classification using recurrent neural networks,” *arXiv preprint arXiv:1705.02636*, 2017.
- [11] R. Pelich et al, “Performance evaluation of sentinel-1 data in SAR ship detection,” in *IEEE IGARSS*, 2015.
- [12] F. Mazzarella et al, “Sar ship detection and self-reporting data fusion based on traffic knowledge,” *IEEE GRSL*, 12(8):1685–1689, 2015.
- [13] European Maritime Safety Agency, “Clean sea net service,” <https://portal.emsa.europa.eu/web/csn>.
- [14] F. Cappello et al, “Grid’5000: a large scale and highly reconfigurable grid experimental testbed,” in *IJHPCA*, 20(4): 481-494, 2006.

MERGING INSAR AND GNSS METEOROLOGY: HOW CAN WE MINE INSAR AND GNSS DATABASES TO EXTRACT AND VISUALIZE INFORMATION ON ATMOSPHERE PROCESSES?

Giovanni Nico (1), Amaia Gil (2), Marco Quartulli (2), Pedro Mateus (3) and Joao Catalao (3)

(1) CNR IAC, Bari, Italy

(2) Vicomtech-IK4, San Sebastian / Donostia, Spain

(3) University of Lisbon, Lisbon, Portugal

ABSTRACT

The use of Numerical Weather Models (NWMs) provides forecasts of meteorological events based on hypothesis on surface cover, boundary layer and other mechanisms of atmosphere processes. In many cases, the selection of appropriate parameters and the tuning of models at the basis of NWMs is made by non-advanced users in *a hoc* manner. The availability of huge databases of InSAR and GNSS measurements of PWV provides those users a unique possibility to falsify the hypothesis at the basis of their choice of parameters. In this work we introduce features, based on statistical analysis and graph theory, that can help to compare InSAR and GNSS measurements to NWM simulations. From the point of view of the user, the meaning of comparison is to quickly catch the informative potential of InSAR and GNSS data if assimilated in a NWM.

Index Terms— Meteorology, Sentinel-1, visualization, databases, Numerical Weather Model (NWM), Synthetic Aperture Radar (SAR), SAR Interferometry, Global Navigation Satellite System (GNSS).

1. INTRODUCTION

Synthetic Aperture Radar (SAR) Interferometry has recently demonstrated its capability to provide useful information about water vapor in atmosphere opening a field of application called SAR meteorology which complement GNSS meteorology [1]-[3]. The knowledge of acquisition geometries, of both SAR interferometry (InSAR) and GNSS data, as well as of the corresponding processing parameters (e.g. the Vienna mapping function for GNSS data) can help to visualize the InSAR and GNSS measurement processes, compute the synthetic data corresponding to the different hypothesis on the description of the atmosphere physical processes and compare with the real measurements [4]-[6]. The launch of both satellites A and B of Sentinel-1 SAR mission of the European Space Agency opened new perspectives in the mapping of Precipitable Water Vapor (PWV) with an unprecedented spatial resolution up to 25 m

and a temporal update of 6 days over the same region [7]. It has been demonstrated that this information can describe meteorological phenomena, such convective processes, with spatial details not captured by dense networks of GNSS sensors. Figure 1 shows an example of availability of GNSS and Sentinel-1 data over the Iberian Peninsula. The temporal sampling of PWV maps provided by Sentinel-1 over a given region can be further reduced if images acquired from different orbits are used as reported in Figure 2.

However, the use of Sentinel-1 data is not so straightforward, and different phase contributions related to both geodetic and atmosphere phenomena must be disentangled. For instance, it should be assumed that terrain displacements contribute with a negligible phase and therefore interferograms are processed using the smallest temporal baseline of six days. Furthermore, SAR interferometry can directly measure only the temporal change of phase propagation delay in atmosphere at the master and slave acquisition times. This delay mainly depends on the turbulence in troposphere (as the laminar component of phase delay in atmosphere is pretty stable in time) and it is canceled out in the interferometric processing. Furthermore, a tiny phase delay due to propagation in ionosphere is also present in Sentinel-1 interferograms also due to its large spatial coverage. In this respect, it can be stated that Sentinel-1 maps of PWV are somehow complementary to GNSS information. In fact, GNSS measurements refer average measurements of propagation delay, both in time (each GNSS estimate refers to a time interval of about half an hour) and space (measurement refers to the average values within the observation cone set during the processing).

The development of tools to visualize information about atmosphere (both troposphere and ionosphere) delay is essential to:

- identify the presence of laminar and turbulence regime in the water vapor spatial distribution
- study the role of terrain morphology and land cover properties on physical processes occurring in the boundary layer

- effectively use the high information content of Sentinel-1 data.

A set of metrics is used to study the temporal correlation of GNSS measurements taken at different locations together with its interaction with

- local topography and land use
- to predict PWV measurements both in space (e.g. at places where no GNSS receivers are available) and in time (near-future time acquisitions).

The relationship between GNSS and InSAR measurement of PWV is studied based on the GNSS and SAR observation geometries and acquisition time intervals. To compute synthetic GNSS and InSAR PWV measurements, NWM numerical simulations are used. The aim of this experiment is to understand how anisotropies in the atmosphere refractivity are mapped in the GNSS processing and how InSAR measurements can help to partially recover those anisotropies. The same metrics developed to study the correlation of GNSS are applied to Sentinel-1 PWV maps, synthetic GNSS PWV profiles and InSAR maps of PWV (or wet delay). In particular, the use of this metrics on synthetic data could be used to identify features and their relationship with the 3D distribution of atmosphere refractivity estimated by NWM output

2. EXAMPLE OF VISUALIZATION OF METEOROLOGICAL INFORMATION

The output of NWMs are basic physical variables, like P, T, relative humidity and so on. These variables are used to compute more advanced magnitudes (e.g. wind distribution, convective available potential energy, etc.) related to the study of the atmosphere. However, these variables can also be used to estimate the 3D distribution of atmosphere refractivity. Figure 3 shows an example of PWV map computed from NWM output using basic physical laws. This map is based on hypothesis at the basis of NWM runs and on numerical solutions of fluid dynamics equations under specific boundary conditions. The comparison of these synthetic PWV maps with the PWV information available in InSAR and GNSS archived data is a way to take benefit from the high spatial resolution data of Sentinel-1 and the real-time GNSS measurements of PWV.

Besides this approach, users in meteorology can benefit from the assimilation of InSAR and GNSS measurements in a NWM [8]. In this case, different scenarios can be studied for atmosphere events and useful thermodynamic quantities. Variables such as the Convective Available Potential Energy (CAPE) and the Convective Inhibition Index (CIN) that are relevant to convective processes in atmosphere as the basis of many extreme weather events can be computed. Figure 4 shows an example of CAPE and CIN maps computed from NWM data over the Iberian Peninsula.

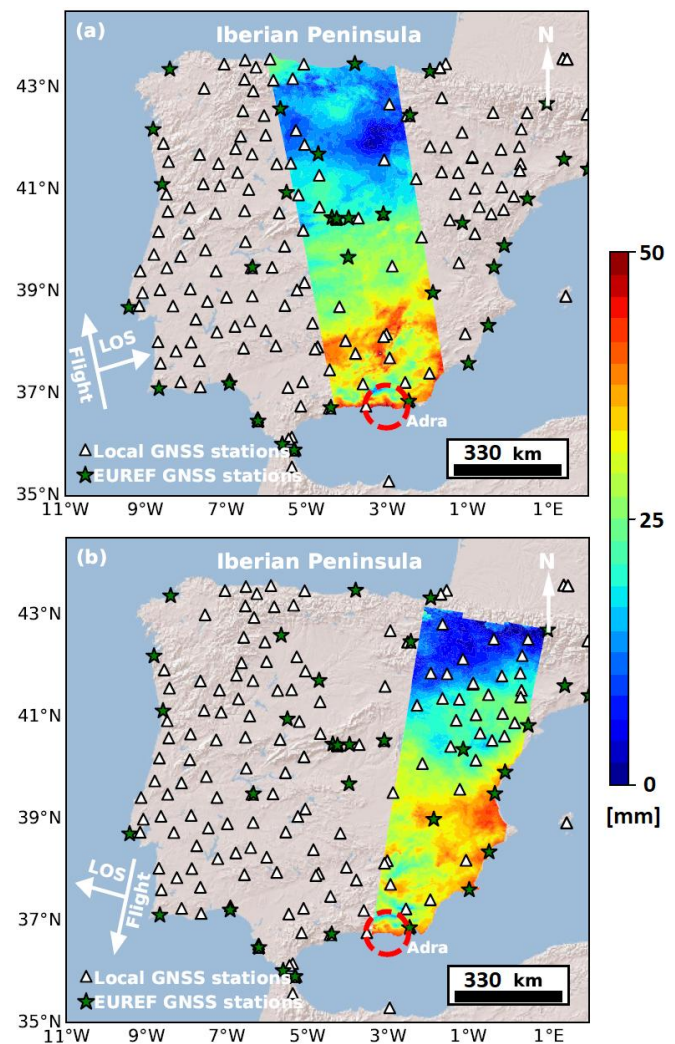


Figure 1: Example of GNSS and Sentinel-1 data acquired over the Iberian Peninsula which are used for the measurement of atmosphere PWV.

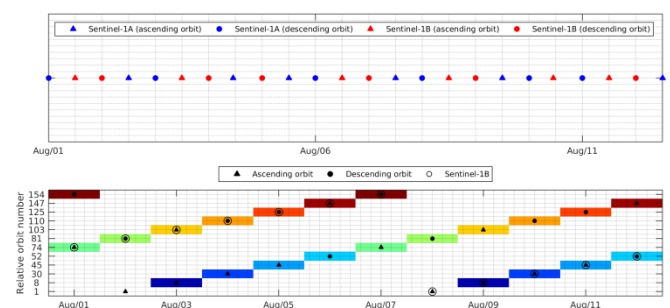


Figure 2: Example of acquisition plane of Sentinel-1 images over the same region. Data are acquired along different orbits, both ascending and descending.

3. STATISTICAL PROCESSING FOR VISUALIZATION

A first step when visualizing the PWV information contained in GNSS and InSAR databases is the comparison between the different data sources and the analysis of the spatial and temporal correlations of PWV data. In order to allow users to follow the evolution of this structure in time we have developed a tool that relies on graph theory to describe the geo-spatial structure of the set of correlation measures among PWV measurements. This approach can be applied directly to GNSS data, where each GNSS station provides a graph node. Figure 5 displays the temporal correlation between GNSS measurement of PWV within a temporal window of 24h. The temporal correlation among GNSS stations can be related to the underlying terrain morphology and land use, as well as the time evolution due to atmospheric phenomena. Figure 6 shows how this temporal correlation is changing in time. The same tool has been applied to study the spatial structure of correlations in Sentinel-1 interferograms and the synthetic PWV maps computed from the NWM output. In this case, a mask has been applied to identify pixels on Sentinel-1 and NWM maps filling the atmosphere portions around each GNSS stations to make a comparison between the different possible data sources.

Visualizations help to evaluate the relationships between different stations and the stability of this relationship in time. This is valuable from both the scientific point of view and for identifying predictive variables to be exploited in modeling efforts. In the case of large databases of direct and indirect measures, visualization options can be particularly relevant in a preliminary exploration phase: the limited availability of methods for the early characterization of extended volumes of data can effectively be complemented by integrating the human visual system in the analysis loop. Specialized techniques [9]-[11] are needed to manage and visually represent extended coverages of remote sensing data in near real-time together with the results of e.g. machine learning regression models allowing a level of interaction that enables visual exploratory statistics on extended datasets.

This work is currently being carried out towards the integration of the described PWV analysis tools with existing frameworks composing web-based mapping interfaces and scalable machine learning engines. The objective of this work is to create a user-oriented tool for the supervised training of classification systems capable of

- detecting extreme situations in the obtained water vapor maps and regression systems
- predicting and smartly interpolating on a global scale the measured values based on ground coverage, geographical and external descriptors
- progressively reducing the deviation of regression results from measured outcomes.

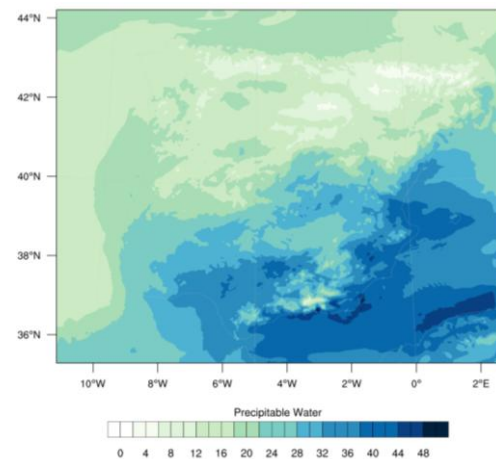


Figure 3: Example of real-time visualization of PWV computed from WRF output

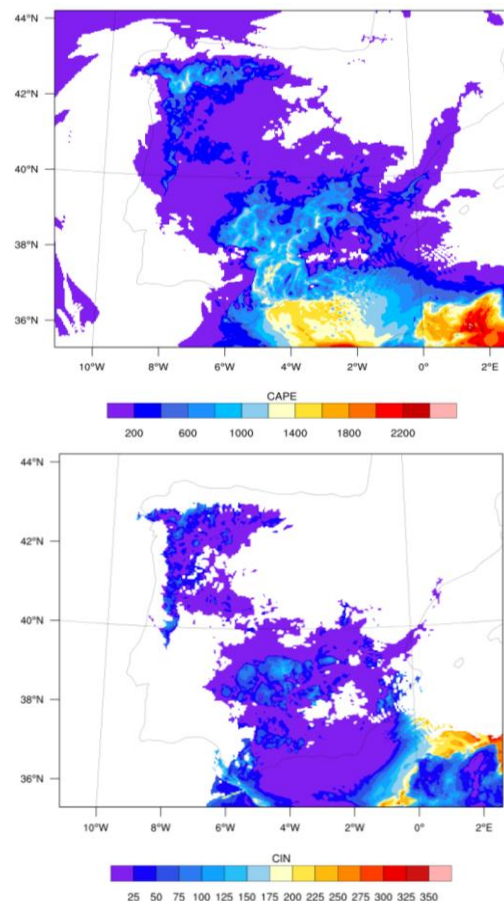


Figure 4: Example of real-time visualization of CAPE (top) and CIN (bottom) information over the Iberian Peninsula.

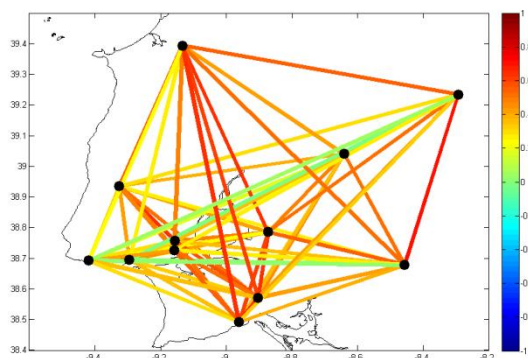


Figure 5: Temporal evolution of correlation between GNSS stations over the western part of the Iberian Peninsula.

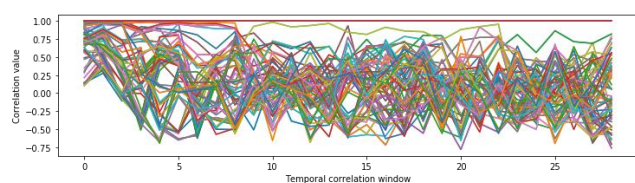


Figure 6: Time series of correlation among GNSS stations over the western part of the Iberian Peninsula.

4. CONCLUSIONS

We have introduced a methodological framework for the exploration of Numerical Weather Models (NWMs) that provides forecasts of meteorological events based on hypothesis on surface cover, boundary layer and other mechanisms of atmosphere processes. The availability of huge databases of InSAR and GNSS measurements of PWV provides non-expert users a unique possibility to falsify the hypothesis at the basis of their choice of parameters. In this work we introduce features, based on statistical analysis and graph theory, that can help to compare InSAR and GNSS measurements to NWM simulations.

5. REFERENCES

- [1] R. Hanssen, Radar Interferometry: Data Interpretation and Error Analysis, Remote Sensing and Digital Image Processing, vol. 2, 1 ed., Springer, Netherlands, 2011.
- [2] Y. Kinoshita, M. Shimada, M. Furuya, InSAR observation and numerical modeling of the water vapor signal during a heavy rain: A case study of the 2008 Seino event, central Japan, *Geophysical Research Letters*, 40(17), 4740–4744, doi:10.1002/grl.50891, 2013.
- [4] P. Mateus, G. Nico, J. Catalão Can spaceborne SAR interferometry be used to study the temporal evolution of PWV?, *Atmospheric Research*, 119, 70–80, doi:http://dx.doi.org/10.1016/j.atmosres.2011.10.002, 2013.
- [4] P. Mateus, G. Nico, R. Tomé, J. Catalão, P. M. A. Miranda, Experimental Study on the Atmospheric Delay Based on GPS, SAR Interferometry, and Numerical Weather Model Data, *IEEE Transactions on Geoscience and Remote Sensing*, 51(1), 6–11, doi:10.1109/TGRS.2012.2200901, 2013.
- [5] P. Mateus, G. Nico, J. Catalão, Uncertainty Assessment of the Estimated Atmospheric Delay Obtained by a Numerical Weather Model (NMW), *IEEE Transactions on Geoscience and Remote Sensing*, 53(12), 6710–6717, doi: 10.1109/TGRS.2015.2446758, 2015.
- [6] P. Mateus, G. Nico, J. Catalão, Maps of PWV Temporal Changes by SAR Interferometry: A Study on the Properties of Atmosphere's Temperature Profiles, *IEEE Geoscience and Remote Sensing Letters*, 11(12), 2065–2069, doi: 10.1109/LGRS.2014.2318993, 2014.
- [7] P. Mateus, J. Catalão, and G. Nico, Sentinel-1 Interferometric SAR Mapping of Precipitable Water Vapor Over a Country-Spanning Area, *IEEE Transactions on Geoscience and Remote Sensing*, 55(5), 2993–2999, doi:10.1109/TGRS.2017.2658342, 2017.
- [8] P. Mateus, R. Tomé, G. Nico, J. Catalão, Three-Dimensional Variational Assimilation of InSAR PWV Using the WRFDA Model, *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7323–7330, doi:10.1109/TGRS.2016.2599219, 2016.
- [9] P. Mateus, G. Nico, J. Catalão, Using TerraSAR-X SAR interferometric data to derive maps of the atmospheric phase delay, *IEEE International Geoscience and Remote Sensing Symposium*, Munich, 3819–3822, 2012
- [10] M. Quartulli, I. Olaizola A review of EO image information mining, *ISPRS Journal of Photogrammetry and Remote Sensing*, 75, 11–28, 2013
- [11] J. Lozano, M. Quartulli, I. Tamayo, M. Laka, I. Olaizola, Visual analytics for built-up area understandign from metric resolution Earth Observation data, *ISPRS International Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, VOL, 1, Issue 2, pages 151–154, 2013
- [12] J. Lozano, N. Aginako, M Quartulli, I Olaizola, E. Zulueta, Web based supervised thematic mapping, *Selected Topics in Applied Earth observation and Remote Sensing, IEEE* , Vol 8, Issue 5, pages 2165–2176 2015

INNOVATIVE APPROACH FOR PMM DATA PROCESSING AND ANALYTICS

*R. De March¹, M. Deffacis², F. Filippi¹, A. Fonti¹, C. Leuzzi¹, M. Montironi³,
A.F. Mulone¹ and R. Messineo¹*

¹ Data Processing and Scientific Applications, ALTEC SpA, Turin, Italy, 10146,
ruben.demarch@altec.space.it, fabio.filippi@altec.space.it, andrea.fonti@altec.space.it, chiara.leuzzi@altec.space.it,
angelo.mulone@altec.space.it, rosario.messineo@altec.space.it

² Mission Operations and Training, ALTEC SpA, Turin, Italy, 10146, maurizio.deffacis@altec.space.it

³ Infrastructure Communication Technologies, ALTEC SpA, Turin, Italy, 10146,
marco.montironi@altec.space.it

ABSTRACT

ALTEC started to work on a framework with the main aim to process a big amount of data allowing a seamless connection between the collected information and the analyses performed by end users. It allows to organize data in the most adapt domain data store in order to have data ready for complex data analysis.

Within this context, the ASDP environment for PMM data was defined and developed. In particular, the PMM module of the ISS is a reference case for the survey on framework capabilities for telemetry data management. The main objective is to demonstrate the advantages achievable through the application of new data analysis methodologies and tools after data organization through ASDP capabilities.

Index Terms— ALTEC, ASDP, ISS, PMM, TECSEL2, Big Data, Data Mining, Data Analytics, Data Processing

1. INTRODUCTION

In the last few years, the integration of innovative approaches for the management of telemetry data showed how the operational activities could be improved. Different research initiatives demonstrated the effectiveness of new techniques in supporting operators working on space missions. In particular, daily operations as well as long-term analyses can be performed with better control and reduction of efforts. These innovative solutions provide all the instruments to enhance the operational process, from telemetry data monitoring to potential anomalies and failures identification. Current technologies show promising capabilities for a wide new set of potential applications and data exploitation strategies.

The main objective is represented by the definition and implementation of a distributed environment where all the resources and analyses can be shared among users and exploited in a more integrated way with respect to the current and traditional approaches.

ALTEC accomplished these goals by building up the ASDP environment.

2. ASDP

ALTEC Space Data Processing (ASDP) is a distributed data processing framework dedicated to the on-ground handling and transformation of any aircraft and spacecraft data. Its modular architecture gives the flexibility for easy adapting to several other domains with commonalities with ground centres, even outside the aerospace domain.

ASDP is not a self-standing solution but it allows integrating both existing and new coded algorithms, enabling automatic processing of large datasets and complex pipelines. This framework is implemented using several state-of-art IT technologies allowing the system to be maintainable, robust, scalable and easily extendible. It enables to organize data in the most suitable domain data store in order to be ready for complex data analysis. Innovative analytics algorithms can be easily activated in order to mine data and extract relevant information.

ASDP takes advantage of containers technologies. This simplifies its deployment in any distributed environment, allows runtime expansion of the system and eases the integration tests. ASDP uses:

- AKKA framework for messaging and cluster managing [1];
- Elasticsearch as metadata repository [2];
- Apache Parquet for file storage of the products [3];
- Apache Cassandra for database storage of the products [4].

One of the ASDP usages is represented by telemetry and other data acquired from the International Space Station (ISS). These data have been ingested and processed in order to extract information, detect and predict anomalies, deduce information patterns useful to operate the systems and to plan maintenance activities. In particular, the PMM Module provides a huge amount of interesting telemetry input data that can be processed with ASDP.

Looking forward, ASDP will process also Exomars' mission data, as ALTEC will host the Rover Operation Control Centre (ROCC) of the mission.

3. PMM

The Permanent Multipurpose Module (PMM), called Leonardo, has been conceived as a logistics module to provide upload cargo to increase the stowage volume at the ISS and to support ISS outfitting, utilization and maintenance as well as its crew logistics. It is derived by the upgrading of the Multi-Purpose Logistic Module (MPLM) Flight Unit 1 (FM1), re-designed to perform its nominal upload logistics mission and then to remain permanently docked to the ISS. PMM has been placed into orbit with the space shuttle mission STS-133/ULF5 on February 24, 2011 and it has operated on the ISS continuously for 80 months. PMM is composed of the following different subsystems:

- Structure & Mechanism (PMM S&M);
- Environment Control and Life Support System (PMM ECLSS);
- Thermal Control Subsystem (PMM TCS);
- Avionic and Power Subsystems (PMM AVS);
- Command & Data Handling (PMM C&DH);
- Software (PMM S/W).

PMM has been designed to operate both in the Shuttle Cargo Bay and at the ISS to provide:

- pressurized and protected habitable environment;
- thermal and environmental control in compliance with standard ISS characteristics;
- avionics functions to operate and control the operational PMM systems;
- sixteen International Standard Payload Rack (ISPR) compatible locations for the accommodation of passive cargo plus additional cargo accommodation capability (bags) in the cone area.

During the On-Orbit Operations phase, PMM provides a habitable environment, even if there is no temperature selectability and PMM dedicated humidity control, but this is in line with requirements, considering that nominal operations are mainly related to the cargo transfer loading and unloading operations [5].

All PMM data are currently stored in ALTEC through the direct link with NASA (ASINet) and analysed by PMM sustaining engineering team. The availability of such a huge amount of data (PMM 2011-2017 telemetry) represents an interesting scenario for the assessment of ASDP data analytics capabilities, and its analysis allows collecting a lot of information on PMM system and equipment, paving the way to apply innovative data analysis techniques.

ASDP is used to support PMM operational activities concurrently with the traditional tools and methodologies. This approach permits to better identify the possible areas of improvements of the framework but at the same time it shows how the traditional processes can be supported and made



Figure 1: PMM docked to the Nadir port of ISS Node 1 during the STS-133 mission.

more effective. Fault prediction techniques have also been tested due to the availability of past data on maintenance activities. The information on anomalies, collected during the mission, provides a useful reference to compare the faults that can be predicted with the real ones.

The presence of data on different subsystems and on equipment characterized by different behaviours represents a well-suited scenario to assess the processing capabilities of ASDP. Innovative analytics techniques help to manage PMM data in a more effective way, improving the current Operations practice of the module [6].

4. DATA ANALYSIS

Two main kinds of data analytics of time series datasets coming from PMM sensors are provided: descriptive analyses and predictive analyses.

4.1. Descriptive analysis

Some basic statistics on some specific datasets are already computed and presented in the reports generated automatically and daily by ASDP. Users can also select a sensor they are interested in and compute, through the ASDP web interface, basic statistics directly as long as complex analyses like trend analysis, outlier analysis and periodicity individuation.

Multi-dataset analyses are present too, like the correlation with delay (between two or more datasets) which aims at finding out if a quantity has an impact over another one and at inferring the time it takes the effect to show itself. This is accomplished by computing the correlation value between a fixed time series dataset and another one that is gradually shifted by a determined time interval. The same approach has been used in TECSEL2, another ALTEC project [7].

Anomaly detection algorithms are also available. Mutual user-defined distances computation, clustering algorithms and noise analysis are some of the techniques applied to

samples of distinct time intervals in order to seek unexpected behaviours.

4.2. Predictive analysis

Advanced machine learning algorithms are available in order to predict future values or to replace unknown (usually missing) values, that could be either measured by sensors or dependent on a group of sensors. Both supervised models and unsupervised models make possible the prediction of a component end-of-life or the explanation of relations among quantities. Therefore, you can obtain information not only on the mere telemetry data, but also on quantities derived from the very observations and on phenomena that may require human intervention.

These methods include Linear Regression as long as Neural Networks, Random Forests and Gradient Boosted Trees. Some members of our team also applied these prediction techniques during the participation in the Mars Express Power Challenge, promoted by ESA in 2016 [8].

5. SPECIFIC ANALYSES ON PMM DATA

5.1. Cabin Fan Assembly (CFA)

While monitoring the speed parameter of PMM CFA, you can observe some spikes, without a constant frequency or size. A spike consists in a speed value greater than the threshold value and with a duration of a few seconds at most.

It is interesting to perform basic statistical analyses on size, frequency, periodicity and duration of these spikes, eventually averaging inside some time windows. In fact, the obtained results help the team to analyze a wide amount of data on a quite long time frame, enabling a deeper understanding of CFA equipment and foreseeing long-term performances.

The CFA telemetry dataset has also been studied in conjunction with fans datasets coming from other ISS modules with the clustering and correlation techniques mentioned above, in order to compare and check the different telemetries. In particular, we want to understand if the behaviour (e.g., average duration and frequency) of the spikes changes from one module to the other.

5.2. Shell heaters

The topic of interest here is the period between the activation and the following shutdown of the chains of the heaters of PMM. There are 19 of these chains, and for every one of them the current, the electrical resistance and the dissipated power are examined, for example by means of a trend analysis.

Moreover, we compute the correlation between different pairs of chains, as long as the correlation with delay between the heaters currents and the cabin air temperature detected inside PMM.

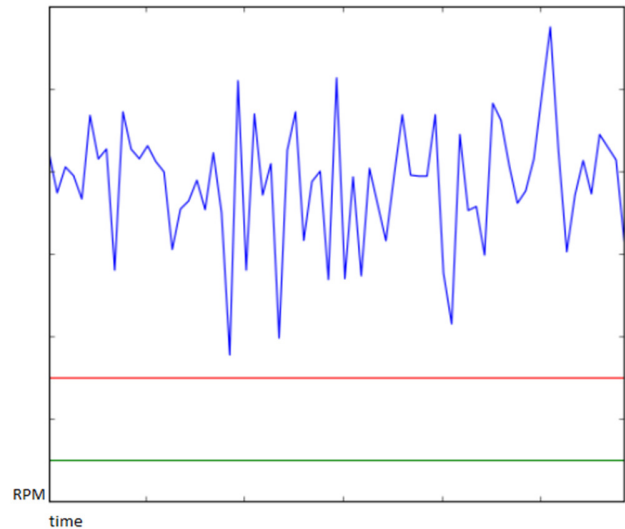


Figure 2: Per-temporal-window-average values of the spikes of CFA speed. The straight lines below indicate the nominal value and the maximum admitted value.

5.3. Cabin pressure

The most illustrative application of the multi-dataset descriptive analyses is the group composed by the three PMM sensors that measure the cabin pressure.

Since the distinct sensors measure the same quantity, the computation of correlations and the anomalies detection inside each one of these datasets are the best techniques to investigate these telemetry data. In fact the aim is to check their homogeneity and consistency as well as how they maintain over time.

5.4. General Light Assembly (GLA)

ASDP descriptive results on hardware performance of PMM GLA, in particular in the proximity of anomalies and unexpected behaviours, led to a better planning of maintenance activities.

Anyway, the main analysis for GLA is not descriptive, but predictive. In fact, the goal of the GLA analysis is predicting when a failure will occur on one of the GLA components. The prediction is based on the trend of the current between the GLA components; in particular, the features present in the GLA telemetry data useful for this prediction are:

- the time the current reaches the alerting threshold value;
- the time the current reaches the failure threshold value after having overcome the alerting threshold one;
- the size of the current spike detected when the GLA is turned on (In-Rush Current).

Moreover, the behaviour of the GLA has not been the same over the years. In fact, during the first two years of PMM activity the GLA were always kept at full-bright level,

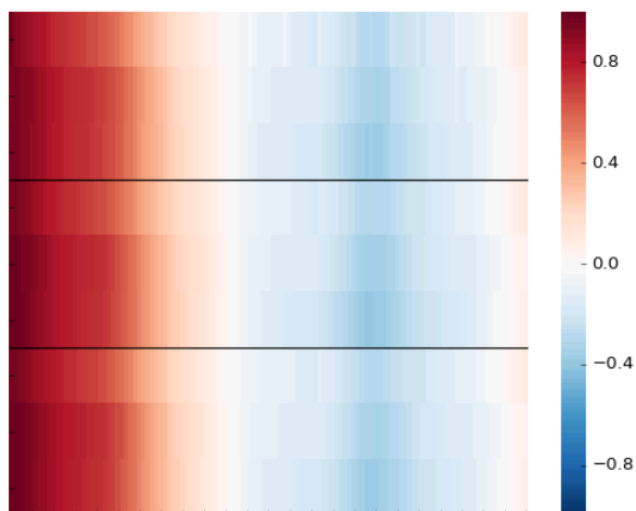


Figure 3: Pairwise correlation with delay between three datasets coming from sensors measuring the PMM cabin pressure. The abscissa axis contains 120 shifted 1-hour time intervals, while the legend indicates the value of the correlation. We can observe how the correlation value is bigger (nearly 1) for little time shifts and decreases when the time shift applied increases, and reaches its minimum, almost a perfect anti-correlation, at about 85 time shifts.

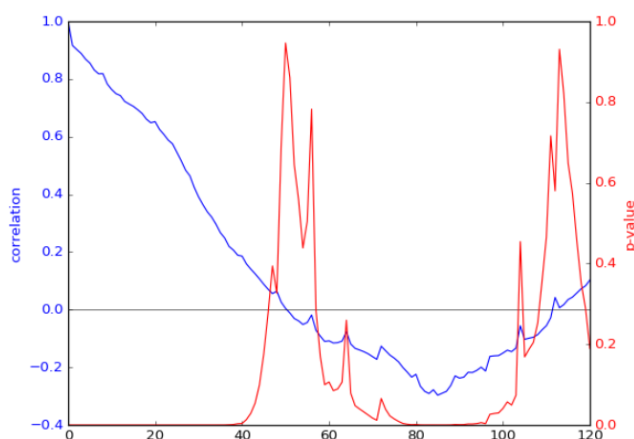


Figure 4: Bi-dimensional version of the second row of Figure 3, which represents the correlation with delay between time series from the first and the second pressure sensor (the latter is the shifted one). In addition, the p-value expressing the significance of the correlation is provided.

while in the following period a dimming strategy has been implemented in order to extend their operative life.

Due to the multiple heterogeneous predictors mentioned above, the best practice here is to apply the random forest algorithm.

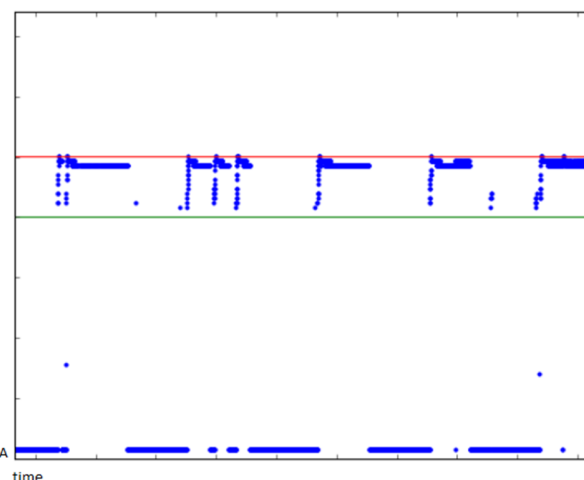


Figure 5: Example time series plot of one GLA current, with the alerting and failure thresholds.

6. CONCLUSIONS

An integrated environment like ASDP can help to keep under control a wide amount of information, showing the operators only relevant data or unexpected situations.

More complex analyses on PMM equipment provide remarkable means to support trouble-shooting, exploiting a big amount of data in a more straightforward way than the classic approaches used until now.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge the Agenzia Spaziale Italiana PMM Program Manager Marino Crisconio for granting the access to the PMM telemetry data.

8. REFERENCES

- [1] Akka, available at <https://github.com/akka/akka> [Accessed 9 October 2017]
- [2] C. Gormley and Z. Tong, *Elasticsearch: The Definitive Guide*, O'Reilly Media, 2015.
- [3] Apache Parquet, available at <https://parquet.apache.org/documentation/latest/> [Accessed 9 October 2017]
- [4] Apache Cassandra, available at <http://cassandra.apache.org/> [Accessed 9 October 2017]
- [5] "Space engineering - Ground systems and operations"; ECSS-E-ST-70C, 31 July 2008; ECSS Secretariat, ESA-ESTEC, Requirements & Standards Division, Noordwijk, The Netherlands.
- [6] "Assessment of an Innovative Data Processing Framework on PMM Space Mission", SpaceOps Workshop 2017, 26-28 June 2017, Moscow, Russia.
- [7] R. De March et al., "TEmporal Characterization of the remote SENSors response to radiation damage in L2", in *Proc. Big Data from Space (BiDS16)*, IEEE, 2016.
- [8] Mars Express Power Challenge, available at <https://kelvins.esa.int/mars-express-power-challenge/home/> [Accessed 9 October 2017]

BENCHMARKING C++ IMAGE PROCESSING LIBRARIES FOR THE EUCLID SCIENCE GROUND SEGMENT

Peter Kettig, Antoine Basset

Centre national d'études spatiales (CNES), Toulouse, France

ABSTRACT

In this paper, we tackle the issue of selecting a C++ image processing library for the Euclid science ground segment (SGS). We propose a new benchmark to objectively compare libraries according to both static, development-related, and dynamic, execution-related, criteria. The latter evaluate the performance of the executable both for single- and multi-thread executions. Furthermore, instead of comparing isolated functions as generally done, we implement a realistic use case: a complete processing pipeline. This makes the results more trustworthy than with classical single-function benchmarks. Eventually, the benchmark will allow the SGS system team to select the most appropriate library.

Index Terms— Euclid, science ground segment, benchmark, C++, image processing

1. INTRODUCTION

The Euclid space telescope is currently under development within ESA's cosmic vision program and set to be launched in 2020. With an estimated amount of 175 PB of data [1] processed by the science ground segment (SGS) over the 6-year mission, thus being representative for the big-data paradigm, it is a key task to process the images efficiently. The SGS is composed of processing functions (PFs), many of which are image processing pipelines implemented in C++, themselves made of smaller pipelines termed processing elements (PEs). In order to guarantee good performance and maintainability of the PEs, an efficient C++ image processing library has therefore to be carefully selected.

Benchmarking libraries often consists in profiling single functions independently to monitor their dynamic behavior in resources and computation time [2, 3]. Yet, state-of-the-art image processing libraries offer much more than just a set of functions: they also help with connecting them as a pipeline [4]. Therefore, we propose to implement a *reference pipeline* with each of the candidate libraries, and compare their dynamic behavior during production with this realistic usage. The *dynamic measurements* are split into single-thread and multi-thread benchmarks. This aims at evaluating the intrinsic parallelization potential of the libraries.

We also introduce *static measurements* on the libraries. They enclose criteria related to the development and maintenance phases to show a developer's point of view. Indeed, development and maintenance times have to be included in the global budget, as we expect the processing methods to evolve even in production phase.

The paper is organized as follows. In the next section, we briefly introduce the reference pipeline to be implemented, and the set of image processing libraries to be compared. In Section 3, we detail the static criteria and report the associated results. Dynamic measurements are presented and their results discussed in Section 4. We finally conclude in Section 5, where we also propose improvements to this work.

2. REFERENCE PIPELINE AND CONTENDERS

To compare the libraries, a reference pipeline is used as a case study. We chose the cosmic ray detection method L.A.Cosmic [5], because it is foreseen as a central component of several PFs for the visible, near-infrared and spectroscopic image processing. It also involves many functions that an image processing library certainly has to cover, such as convolution, morphological or arithmetic operations.

The following 11 contenders were found suitable for being benchmarked: CImg, CxImage, DevIL, ExactImage, GIL, ITK, Magic++, MKL, OpenCV, Simd, and Skia. Yet, implementing the algorithm for so many contenders is too time-consuming and sometimes not even possible due to the lack of functions available. For that matter, let us introduce the static measurements.

3. STATIC MEASUREMENTS

Static measurements give information about the comprehensiveness and compatibility from a developer's point of view. They are the following:

License According to the SGS rules, the license of the library must be LGPL-compatible.

Completeness The library has to contain as many functions as possible to implement the reference pipeline; otherwise the developers will have to implement missing functions himself, at the expense of development time and performance. The measurement is the number of

Table 1: Static measurements

Library	License	Complete	Doc.	Lines of code
Clmg	yes	100%	92%	835
CxImage	yes	88%	-	-
DevIL	yes	83%	-	-
ExactImage	yes	72%	-	-
GIL	yes	78%	-	-
ITK	yes	100%	97%	1540
Magic++	yes	83%	-	-
MKL	no	83%	-	-
OpenCV	yes	94%	94%	1300
Simd	yes	78%	-	-
Skia	yes	67%	-	-

available functions divided by the number of needed functions required by L.A.Cosmic.

Documentation The functions should be clearly documented. The measurement is the ratio of fully documented functions over available ones.

As reported in Table 1, among the eleven contenders, only three offer a completeness of more than 90%: Clmg, ITK, and OpenCV – GIL and CxImage are close to that number but lack the median filter, which is prohibitive. They are thus used to implement L.A.Cosmic, and compared to the existing Python implementation¹ used as a scientific reference.

During the implementation, data is collected about missing documentation, thus appending the list of static measurements for the three finalists, which is also shown in Table 1. Although subjective, the number of lines of code gives a good indication on the implementation complexity.

4. DYNAMIC MEASUREMENTS

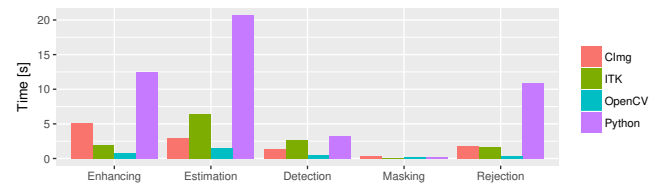
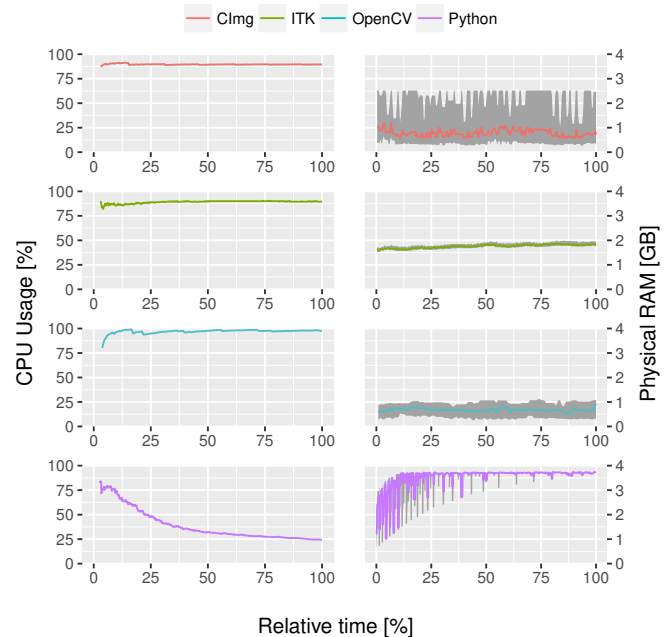
The dynamic measurements we perform consist in monitoring the CPU and RAM usage, as well as the computation time of each implementation. To measure the resource usage, we use statistical profiling which is preferred over other methods [3] as the goal is to run the executable as close as possible to future production setup.

We split the reference algorithm into five steps, each of them focusing on a specific type of image processing [5]: cosmic ray enhancing (convolution, up- and downsampling), background noise estimation (arithmetic operations), cosmic ray detection (non-linear filters), masking of cosmic rays (binary operations), and rejection of saturated stars (morphological operations).

4.1. Single-thread benchmark

All of the implementations are run in the exact same configuration as it would be set up in production by using the Euclid

¹Python implementation 0.4 of L.A.Cosmic, http://obswww.unige.ch/?tewes/cosmics_dot_py/

**Fig. 1:** Computation times for the five algorithm steps**Fig. 2:** CPU and RAM usage for each contender over 36 sequential runs (at 0%, first run starts; at 100%, last run ends)

reference virtual machine [6] and simulated images. Smallest images have 1 kpx while the biggest ones correspond to the full Euclid CCD size of 16 Mpx. Resources are fixed to 1 CPU, 4 GB of physical memory and 8 GB of virtual memory.

As reported in Fig. 1, the reference Python implementation is the slowest of the four. OpenCV, on the other hand, performs globally better than the other libraries, although not for the masking step, where ITK is the fastest.

Fig. 2 shows that, when running L.A.Cosmic thirty-six times in a row (the Euclid visible imager is made of 36 CCDs), Python suffers the most from the restricted resources, leading to a situation where the program cannot proceed due to the lack of memory. On the other hand, OpenCV is not only faster than the other contenders but also more resource-efficient. The memory consumption of ITK is not stable over the runs; Further investigations exhibited a memory leak in the library.

While the comparison of Python vs. the C++ contenders could offer an interesting comparison of two languages, it is not done on an equal basis. Indeed, the Python version is an

Table 2: AIC for each sizing-model and each contender

Sizing model	CImg	ITK	OpenCV
$O(n)$	-1095.90	-553.83	-1416.82
$O(n \log n)$	-848.75	-111.39	-1149.97
$O(n^2)$	41.99	733.22	-327.95
$O(n^3)$	218.68	905.84	-154.66

off-the-shelf component which should be refactored for fair comparison. In the following, the focus is thus solely on the C++ implementations.

To evaluate the computation time dependency to the image size, termed sizing, the AIC [7] was used to compare existing models. The samples consisted of 200 images with linearly increasing size n from 1 kpx to 16 Mpx, and multiple runs. Classic sizing models $O(n)$, $O(n \log n)$, $O(n^2)$ and $O(n^3)$ were tested. Table 2 shows that the linear dependency is the best fit for all libraries.

4.2. Multi-thread benchmark

Parallelization is sometimes necessary to efficiently process the data. Multiple ways to parallelize exist, reaching up to the level where the whole pipeline is run in parallel. In this paper, we evaluate the parallelization at the image level, meaning that the library has to take care of splitting up the input image, distributing the workload evenly on all threads and finally putting the produced outputs back together again. This means only using a parallelized version of the function and not parallelizing the algorithm itself.

All C++ contenders offer the possibility to rely on OpenMP or pThreads by simply setting specific flags [8]. The Intel TBB library can be used by OpenCV as well but the license is not LGPL-compatible. The results are produced by increasing the number of threads from 1 to 8, with 8 GB of physical and virtual memory.

Figure 3 shows that ITK scales well with the increasing availability of processing power. OpenCV and CImg suffer heavily from false sharing, which is the problem of a thread unwillingly stopping others from proceeding with their program execution due to the reservation of independent variables residing in the same processor cache. This is a result of a processor not reserving a single variable but rather a whole cache-line.

In order to avoid race-conditions [9], the other process has to wait until those variables are free again. This slows down the processing even further than the single-thread benchmark and is reflected in the CPU usage for each algorithm: Looking at the 4-CPU measurement in Fig. 4, ITK is using almost all available resources whereas OpenCV and CImg use only one of the available four threads.

Solving the false sharing problem is non-trivial: A simple approach is to declare all shared resources constant, and

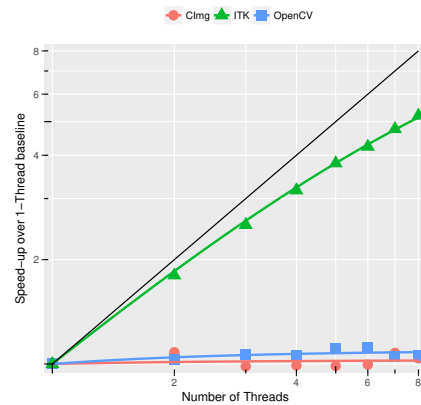


Fig. 3: Speed-up factor versus the number of threads (marks) with Amdahl's law fitting (lines)

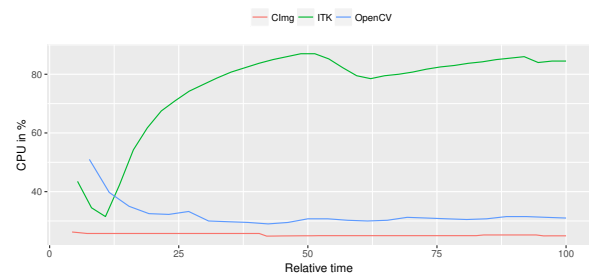


Fig. 4: CPU usage for 4 cores

therefore read-only. This can be the solution for some functions, but not for in-place ones. Here, the rate of access to the shared memory has to be reduced, by using padding around this memory as well as alignment if possible. However, this compromises the image container concept presented by both OpenCV and CImg, making a completely new version necessary, as the whole memory layout would have to be revised.

To model the behavior of the multi-thread implementations, three models are compared: the linear and logarithmic ones, as well as Amdahl's law [10]. The latter assumes that using twice the amount of resources will not result in half of the execution time but is rather dependent on the portion of the code that benefits from the improved resources. Thus, parallel computing is only reasonable for programs with a high portion of parallelized code, otherwise the resources will be wasted.

The Amdahl's law writes:

$$S(f, s) = \frac{1}{1 - f + \frac{f}{s}} \tag{1}$$

with S the speed-up, f the portion of the task that benefits from multi-thread processing and s the speed-up of the former part – the number of cores to a first approximation.

In Table 3, we compare the aforementioned scaling models by using AICc [11]; The correction is necessary due to the

Table 3: AICc for each scaling model and each contender and estimated portion \hat{f} of parallel code in Amdahl's law fitting

Scaling model	CImg	ITK	OpenCV
$S(s) = s$	16.58	10.83	16.89
$S(s) = \log s$	16.35	13.38	16.37
Amdahl's law	-23.28	-19.02	-30.64
with \hat{f}	2.6%	92.0%	8.6%

small sample size of only eight multi-threading levels (from 1 to 8) and two runs each. The table shows that Amdahl's law is always the best fit, which allows us to plot this curve in Fig. 3. By fitting the Amdahl's law, an estimate \hat{f} of the portion of parallel code is obtained, which is also reported in Table 3.

ITK is performing best in this benchmark. However, the result is also dependent on the design of the executable, which is up to the programmer. Indeed, we only set here the parallel flags of the libraries, while better results would be obtained by parallelizing at a higher level, i.e., the algorithm itself. Still, what the table shows is that ITK is internally better parallelized than the other contenders.

5. CONCLUSION AND FUTURE WORK

We have proposed a new way of benchmarking libraries in two ways. First, by introducing static criteria, which are development- and maintenance-related measurements – two phases which should certainly not be neglected when selecting a library. Secondly, a full real pipeline was implemented with selected contenders instead of comparing independent functions. With such a more realistic use case, results are simply more trustworthy.

The static criteria also helped to sort out the library contenders reliably, and find three suitable candidates (CImg, ITK and OpenCV) even without an implementation. The documentation measurement reflects the user friendliness of each candidate, with ITK being the easiest to begin developing with.

The dynamic measurements were conducted on a system as close as possible to the final configuration, and with Euclid-like images. This realistic integration together with a split of the algorithm into coherent computation steps gives results that are more comprehensive than single-function benchmarks. The three implemented contenders showed similar performances in terms of computation time for the single-thread benchmark; they were also much faster than the reference Python implementation. On the memory side, OpenCV is more sparing than the other implementations.

Finally, the multi-thread benchmark showed that ITK is best suited for multi-threaded processing with the given implementation. The single- versus multi-thread issue is the final step to select a contender: Choosing a level of parallelism at the image-level gives a clear recommendation in fa-

vor of ITK, whereas the pipeline-level parallelism would favor OpenCV and CImg.

In the future, the list of functions will be extended by implementing a second algorithm. The static criteria could be extended with additional information on the the library maintenance, such as bug fixing rate. Last but not least the whole benchmark could be extended by using other libraries, or even other programming languages. Halide [4], for example, features high-level optimizations of pipelines and therefore offers a high potential; It is also compatible with C++.

6. REFERENCES

- [1] M. Poncet, C. Dabin, J.-J. Metge, K. Noddle, M. Hollimann, M. Melchior, A. Belikov, and J. Koppenhoeffler, "Euclid: "Big data from dark space" – Science ground segment challenges for next decade," in *BiDS*, 2014.
- [2] T. L. Falch and A. C. Elster, "ImageCL: An image processing language for performance portability on heterogeneous systems," in *IEEE HPSC*, 2016.
- [3] S. L. Graham, P. B. Kessler, and M. K. Mckusick, "Gprof: A call graph execution profiler," *ACM SIGPLAN Notices*, vol. 17, no. 6, pp. 120–126, 1982.
- [4] J. Ragan-Kelley, C. Barnes, A. Adams, S. Paris, F. Durand, and S. Amarasinghe, "Halide: A language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines," *ACM SIGPLAN Notices*, vol. 48, no. 6, pp. 519–530, 2013.
- [5] P. G. van Dokkum, "Cosmic-ray rejection by Laplacian edge detection," *Publications of the Astronomical Society of the Pacific*, vol. 113, no. 789, pp. 1420, 2001.
- [6] M. Poncet, Q. Le Boulc'h, and M. S. Hollimann, "Euclid: Using CernVM-FS to deploy Euclid processing software on Computing Centres," in *ASP ADASS XXVI*, 2016.
- [7] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
- [8] D. Novillo, "OpenMP and automatic parallelization in GCC," in *GCC Developers Summit*, 2006.
- [9] T. H. Cormen, *Introduction to Algorithms*, MIT press, 2009.
- [10] J. L. Gustafson, "Reevaluating Amdahl's law," *Communications of the ACM*, vol. 31, no. 5, pp. 532–533, 1988.
- [11] S. Hu, "Akaike information criterion," *Center for Research in Scientific Computation*, vol. 93, 2007.

A FRAMEWORK FOR OBJECT DETECTION IN SATELLITE IMAGES

Mathias Ortner¹, Pierre Blanc-Paques¹, Ségolène Bourrienne¹, Laurent Gabet, Jean-François Faudi

Datalab & Image chain Department, Airbus Defence and Space, Toulouse

Abstract

High-resolution satellite imagery has become widely available, enabling numerous applications. One of the challenges for these applications is the extraction of semantic information from the images: it is now almost impossible for human operators to cope with the increasing volume of data. In this paper, we present a framework which aims at addressing the particular problem of automatic object detection. We propose a full process to access the data, perform training of state-of-the-art detection algorithms, and deploy the algorithms in an operational environment.

Index Terms— Satellite imagery, object detection

1. INTRODUCTION

In the last few years, the increasing volume and resolution of satellite imagery [1] has enabled numerous applications in agriculture, mapping, defense and security. While, a few years ago, the difficulty was to get access to the data, the key challenge is now to understand and index the data: namely to accurately extract semantic information. Automatic object detection partly addresses this challenge by pinpointing objects of interest in the images.

Recent advances in machine learning and deep-learning are tackling the problem of object detection, both in classical photographs and in aerial imagery [2], [3], [4]: state of the art results are now comparable to or better than human performances. From an industrial point of view, the goal is thus not so much to improve the results, but to benefit from these breakthroughs by efficiently integrating the existing technology into operational systems.

2. ONEATLAS

OneAtlas is the commercial name of Airbus Defence and Space digital platform for Earth-Observation imagery. It covers the storage (on line image archive), the processing functionalities, the services and the analytics. OneAtlas offers real-time access to *all* Airbus Defence and Space imagery products (4 optical satellites, 2 radar satellites). The platform also exposes dedicated APIs to programmatically access the data and services.

¹ computervision@airbus.com

By providing image database of unprecedented size (tens of petabytes), OneAtlas is a key element for the development of high-performance machine learning algorithms. In the frame of deep learning for object detection, we rely on OneAtlas to get real-time access to fresh and archive images and efficiently build the datasets used for algorithms training and evaluation.

3. ACTIVE LEARNING

The performances of detection algorithms are conditioned by the availability of a large amount of labeled training images. The annotation process can be difficult, tedious and costly: crowd-sourcing platforms [5] can be used to have humans perform this task at low-cost, but the resulting labels are neither accurate nor reliable. Another way is to use active learning.

Active learning [6] is a successful method for reducing cost and improving labeling accuracy in a classification setting with a large amount of unlabeled data. This is done by:

- starting the learning process with only a handful of labeled examples,
- training a detector on this training set,
- running the detector on unlabeled examples,
- finding the most informative unlabeled examples and asking an operator for their label,
- updating the training-set with the new labeled examples and retraining the detector.



Figure 1: Initial training set (planes and counter examples)

Figure 1 and Figure 2 illustrate the evolution of the training set for the active learning in the example of planes

detection: initially (Figure 1), only single planes are labeled as example (bottom row), and the counterexamples (top row) are random patches. After training a first detector, the training set is then refined using the results (Figure 2) to include difficult counterexamples (top row) and examples of planes (bottom row) in various configurations.



Figure 2: Refined training set

We have developed a dedicated software suite for the active learning loop; including tools for model configuration, training, review of results, update of the datasets. These tools enable to rapidly address various kinds of detection problems, and to flawlessly train state-of-the-art models for dedicated applications.

4. USE CASES AND PERFORMANCES

The main problems which have been addressed are vehicle detection (aircraft, ships, cars) and change detection. Starting from a first training set of a few hundreds of labeled examples / counter-examples, we rely on classical data augmentation techniques (image rotation, translation, contrast adjustment) to produce a dataset of a few thousand examples. The creation of the dataset for the change detection consists in manually tagging areas of interest (change) in pairs of images, keeping in mind that what is “of interest” can be application dependent: while some users may require to carefully monitor the state of crops, some others may consider as false alarms the detection of seasonal changes on fields and agricultural lands.

This dataset is then used to train a first detector. In only 4 to 5 iterations of the active learning loop, and by manually correcting a few dozens of examples at each steps, the algorithms achieve reasonably good performances (Table 1). We use Tensorflow, which enables to deploy the training on multiple machines, with possibly multiple GPUs: with this setup the computation time for the training is about 6 to 8 hours. Various scenes and conditions are chosen during the learning loop to provide the algorithm with the widest possible set of examples.

Table 1: Performance of detection algorithms

	Recall	Accuracy
Aircraft detection	95%	95%
Ship detection	93%	92%
Vehicle detection	94%	70%

The run-time performance depends of course on the size of the input image and the number of machines used for the inference: the problem being completely parallelizable, the load can be balanced on multiple machines by tiling the input. As an order of magnitude, the inference takes about:

- 10 seconds for a 1024x1024 input image on a single CPU for the object detection
- 30 seconds for two 1024*1024 input images on a single CPU for change detection

5. ALGORITHMS INTEGRATION AND DEPLOYMENT

At the end of the training loop, the resulting model is encapsulated in a docker container [7], to be finally exposed as a service in OneAtlas platform (see Figure 3). The service API can then be used to process any image available in the archive. This development and integration process results in unprecedented reactivity to add cutting-edge functions to the operational system.

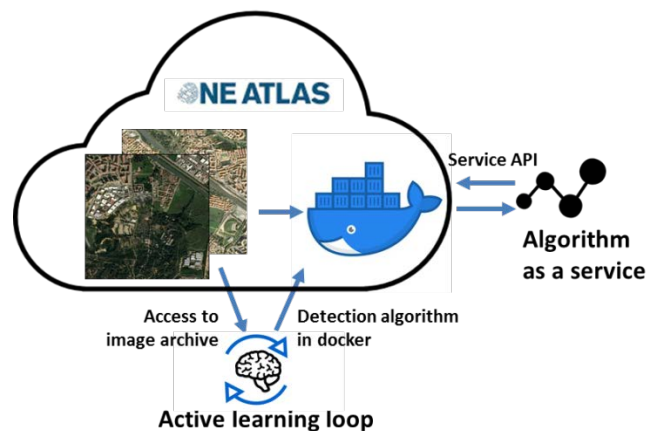


Figure 3: End-to-end framework for algorithm development and deployment

6. CONCLUSION

We have briefly presented a framework covering access to the data, labeling, training of detection algorithms, and deployment. This framework is being used for several applications: detection of aircrafts, ships, vehicles, change detection; and is sufficiently generic to address various other problems. This end-to-end approach enables us to quickly train and deploy state-of-the-art algorithms, thus benefiting from latest breakthroughs in deep learning to add functions and upgrades to operational systems.



Figure 4: Example of Aircraft detection

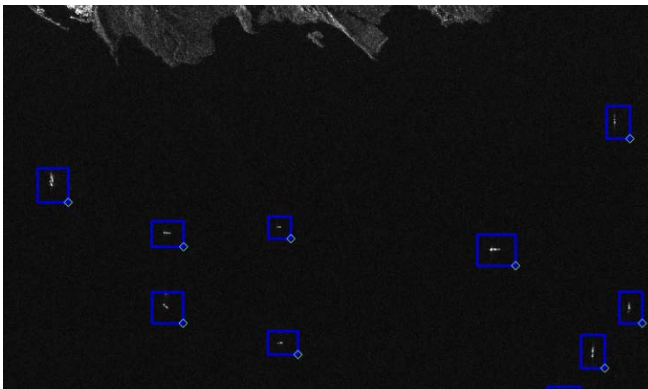


Figure 5: Example of Ship detection



Figure 6: Example of change detection

7. REFERENCES

- [1] Doe, J., “Commercial satellite imaging market - global industry analysis, size, share, growth, trends, and forecast, 2013 – 2019”, *Transparency Market Research*, 2014,
- [2] Ren, S, He, K, Girshick, R, Sun, J, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, *Arxiv*, 2016
- [3] Liu, W, Anguelov, D, Erhan, D, Szegedy, C, Reed, S, Fu, C, Berg, A, “SSD: Single Shot MultiBox Detector”, *Arxiv*, 2016
- [4] Sommer, L, Schuchert, T, Beyerer, J, “Deep learning based multi-category object detection in aerial images”, *Automatic Target Recognition XXVII*, 2017
- [6] Von Ahn, L, Dabbish, L, “Labeling images with a computer game”, *CHI*, 2004.
- [5] Bietti, A, “Active Learning for Object Detection on Satellite Images”, *CalTech*, 2012
- [6] Docker, “Modern Application Architecture for the Enterprise”, *White Paper*, 2016

SCOUTER: GEO-SOCIAL AND REAL-TIME ANOMALY CONTEXTUALIZATION

Badre Belabess^{#△}, Musab Bairat[#], Jérémy Lhez[△], Olivier Curé[△], Houda Khrouf[#], Gabriel Kepekian[#]

[#]Innovation Lab, ATOS F-95870, Bezons, France

[△]LIGM (UMR 8049) F-77454, Marne-la-Vallée, France

ABSTRACT

Anomaly detection is a key feature of applications processing Internet Of Things observations. While in fact, detected anomalies do not necessarily mean abnormal reality situations, external visibility could assist decision making by contextualizing what is happening. Exploiting space domain and social web is the way to build enriched contexts on sensor data. These spatio-temporal contexts hence become an integral component of the anomaly detection and need to be processed using a Big Data streaming approach. In this paper, we introduce Scouter, a generic tool that helps in capturing, analyzing, scoring and storing the contextual information of a given application domain. The processing depends on a semantic-based approach that exploits ontologies to score the relevancy of contextual information.

Index Terms— Big Data, Stream processing, Anomaly Detection, Geo-profiling, social media.

1. INTRODUCTION

Big data and stream processing are nowadays a big hype. Large amounts of data, which are generated every day by streaming resources obtained from the Internet of Things (IoT), are continuously accumulated and stored in Big Data platforms. The analysis of these data supports intelligent software functionalities usually based on machine learning and semantic-based methods. Among these features, anomaly detection [1] is predominant and is tackling domains as diverse as medicine (*e.g.*, to identify malignant tumors in MRI images), finance (*e.g.*, to discover cases of credit card transaction frauds), information technology (*e.g.*, to detect hacking situations in computer networks). In the Waves project¹, we are interested in detecting anomalies in a large network (over 100.000 km) distributing drinkable water to over 12 million clients. Based on real observations, we found out that anomaly detection is highly correlated to the surrounding context. For instance, abnormal high pressure and flow measures are typical of a water leak. But consider these values are measured on a week-end day for a given residential area. Then, high water consumptions could be explained

by a sport or cultural event happening in that area or simply by some hot weather conditions implying garden watering. Hence, an efficient approach for anomaly detection needs a spatio-temporal contextualization including, for example, real-time social events, geographical profiles, weather observations, etc. The tasks related to the capturing, analyzing, scoring and storing of contextual metadata are demanding since they require interactions between software components especially in a distributed environment. These components generally handle stream, natural language and ontology processing as well as the storing of complex data structures, etc. The Scouter tool is open-source and proposes implementations of these tasks by taking care of installation and configuration aspects. Designed as a generic tool, Scouter relies on the declaration of the data sources one needs to retrieve contextual information from, *e.g.*, tweets, POIs, RSS feeds. Scouter core mainly consists of two modules: social media analytics and geo-profiling. The analytics module process the coming social media feeds in terms of Natural Language Processing (NLP) topic extraction and sentiment analysis, which results in a score for each feed. This scoring is based on ontological hierarchy of concepts and represents the relevancy of the feed to the target use case. The geo-profiling module uses OpenStreetMap to build the geographical profile of the area of interest, such as being residential and industrial zone with certain probability. Both the social feeds and geo-profiles will be provided to the end-user to contextualize the detected anomaly. To the best of our knowledge, Scouter is a first attempt at providing a semantic-based and stream processing tool for anomaly contextualization.

2. ARCHITECTURE

Scouter, as component of Waves platform [2], was developed to be generic, fully configurable and easy to use. As a summary, data is fetched from different web sources, processed and stored in a NoSQL database to be fed to a messaging broker for integration with other systems. The main components of our system are: a set of Web data connectors, a social media analytics and geo-profiling units, a messaging broker and a web service. Figure 1 depicts the Scouter architecture.

The web connectors unit consumes data from different data sources: Twitter stream feed, Facebook graph API, RSS

¹<https://www.waves-rsp.org/>

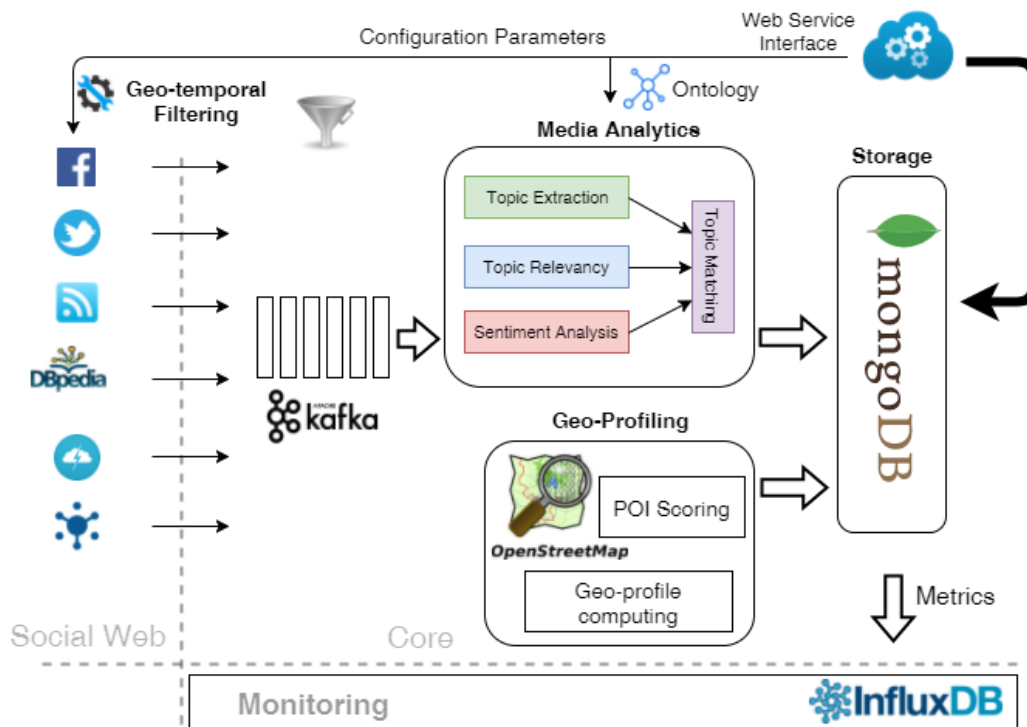


Fig. 1. Scouter Architecture

feeds for newspapers, Open Weather Map, Open Agenda as a source of events. The crawling is executed in an efficient multi-threading mechanism using rest APIs or HTTP connections. Data are fetched and fed to Kafka queue based on: (i) geo-temporal filtering, meaning where and when anomalies have been detected; (ii) keywords which are represented by ontology concepts. It is very important to have a distributed middleware messaging that is efficient in terms of speed, fault tolerance and scalability, and Kafka meets all of these requirements.

The social media analytics unit digests the fetched feeds from Kafka and leverages the distributed Spark framework to analyze feeds in real-time. Feeds are processed to determine relevancy to the use case. For this purpose, sentiment analysis and topic extraction algorithms are used based on Apache OpenNLP library. The sentiment analysis classifies the feed into positive or negative category. The topic extraction parses the text of feed to discover occurrences of terms. Then, the scoring module takes advantage of user defined weights, *i.e.*, a real value in the $[0, 1]$ range, associated to ontology concepts to provide an overall scoring for each text. After scoring, events are recorded into a MongoDB database. Both Spark and MongoDB meets scalability, fault tolerance and speed requirements. As a result, we obtain in real-time spatio-temporal and scored contexts that can assist the end-user to confirm an anomaly detected from IoT observations.

The geo-profiling component aims to determine the geo-

graphical profile for the anomaly area. Using the geographical characteristics (*e.g.*, POI, terrain areas) fetched from OpenStreetMap, a profile is generated that includes different categories that the zone has such as residential, touristic, industrial and agricultural.

Scouter also provides metrics monitoring component to track the performance of the system. Different metrics including query times, feed processing time and topic extraction training time are stored in InfluxDB, which is a time series database with very high reading and writing speed. The REST Web services component is used to be able to configure the system in an easy and human readable way through a user interface.

3. MEDIA ANALYTICS AND NATURAL LANGUAGE PROCESSING

In this section, we detail the methodology of collecting data from the various types of sources available on the web. This process is built on four major components that work together in order to extract the most relevant events, identify the appropriate summaries and avoid duplicate events stored in the database.

3.1. Ontology

Every scrapping tool relies on a configuration file that lists the properties of words, concepts or events that it tries to fetch

[3]. In Scouter's case, the fetching capabilities rely heavily on a pre-built ontology that lists the main concepts the user is looking for. By combining the concepts and the properties through predicates, we can build a complex graph used for the water leak use case. We can easily argue that this type of structure holds more expressiveness than a classic list of keywords exposed in a configuration file.

3.2. Topic Extraction

After fetching the proper events from the various sources based on the ontology of concepts and subconcepts, the next step is to extract meaningful topics from the events. The main pre-processing here concerns the cleaning of the input text, the identification of potential candidates, and finally the stemming and case-folding of the phrases. Input files are filtered to regularize the text and determine initial phrase boundaries, then the splitting into tokens alongside several modifications are made (apostrophes are removed, hyphenated words are split in two, etc). Next, we consider all the subsequences in order to determine the ones that are suitable candidate phrases. To increase the accuracy, we use a list of french stop-word list containing more than 500 words in different syntactic classes (conjunctions, articles, particles, etc). Then we case-fold all words and stem them using the iterated Lovins method [4] to discard any suffix, and repeating the process until there is no further change. Stemming and case-folding allow us to treat different variations on a phrase as the same thing

3.3. Topic Relevancy

Several research works tackle the issue of automatic summarization [5]. The tools proposed generally mix several classes of features such as summary likelihood, use of topic signatures or syntactic analysis [6]. In our case, we chose a direct approach based on distributional similarity that compares input and summary content. In fact, we consider that a good summary should be characterized by low divergence between probability distributions of words in the input and summary, and by high similarity with the input. For this purpose, we used two common measures: the Kullback Leibler divergence and the Jensen Shannon divergence. First, words in both input and summary are stemmed and separated before any computation. Then we compute the two measures:

Kullback Leibler (KL) divergence: It corresponds to the average number of bits wasted by coding samples belonging to P using another distribution Q, an approximate of P. It is given by:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}.$$

Jensen Shannon (JS) divergence: This one leverages on the fact that the distance between two distributions cannot be

very different from the average of distances of their mean distribution. It is given by the following formula:

$$JSD(P || Q) = \frac{1}{2}D(P || M) + \frac{1}{2}D(Q || M) \quad (1)$$

$$\text{where } M = \frac{1}{2}(P + Q) \quad (2)$$

The final step is to use the output of these two functions to rank the extracted topics and keep only the ones with the best summarization score (*i.e.*, lowest divergences).

3.4. Sentiment Analysis

During the last decade, sentiment analysis has known an exponential development due to the growing usage of social networks and the popularity of websites where people can state their opinion on different products and rate them. Many solutions have been proposed and packaged in several technologies [7], we propose in this section a simple approach based on some tools provided by the Stanford CoreNLP toolkit [8]. Since our use case is to analyze media and social networks within the French territory, we used a French dictionary embedded in a wrapper to analyze the words. After the pre-processing phase, we enter the main computation step where we apply the model. Among several models, we chose the compositional one over trees using deep learning. It relies on nodes of a binarized tree of each sentence, including, in particular, the root node of each sentence, are given a sentiment score. In order to capture the sentiment of an input text, a Recursive Neural Tensor Network model (RNTN) is built based on the characteristics of the input phrases.

3.5. Topic Matching

Scouter tries to avoid duplicate events that refer to the same happening or occurrence. In order to complete this task with high accuracy, we are mixing several approaches relying on NLP processing from ontology building to sentiment analysis. The process follows the following steps: For each event fetched from the different sources, the topic extraction phase will propose a list of potential summaries based on a Bayesian approach. Then these summaries will be ranked using the lowest divergences (*i.e.*, KL divergence and JS divergence) in order to assess their accuracy. Among the highest ranked ones, we will check if they have the same sentiment (*i.e.*, positive, neutral or negative). If one of the selected topics during this process have the same sentiment, we assume then that they are referring to the same event in the same way.

4. GEO-PROFILING

The goal of Scouter is to provide relevant information to contextualize and potentially explain detected anomalies. The media analytics module helped in filtering the relevant events

and removing duplicates, however geographical information about the area where the anomaly is spotted could further fine-tune and improve the accuracy of the context. This task is handled by two complementary modules: Geo-profiling module and reasoning module. It can be performed before the reasoning, to orientate the research of events, or after, to change the ranking of the potential sources. It is composed of two methods, that are combined and enriched with a third consumption-based method for better results.

Method 1: It extracts points of interest (POI) present in a given sector, from online geographic data sources, to determine the proportion of different types of surfaces composing it. The domain field expert selected the types of surfaces relevant to our use cases, giving us five profiling parameters: residential, natural, agricultural, industrial and touristic. Then, we created a rating file, assigning notes to each POI, in order to compute a score for each type of surface, and calculate its proportion (*i.e.*, a real value in the [0,1] range).

Method 2: It is also based on geographic data, but uses features modeled as polygons instead of POI. The inclusion tests are more complete, since some polygons may be included completely or partially inside the consumption sector. Also, the computation is not performed using the rating system, but the areas of the polygons, which are less arbitrary. Otherwise, both methods produce the same result: proportions (also as real value in the [0,1] range) for each type of surface.

Method 3: For certain types of consumption sectors, our two methods can slightly differ. To decide which method should be used in each case, we added a third method that computes what we denote as the consumption ratio. For each sector, we compute the daily flow, and make an average over a long period of time to avoid anomalies; then we divide this flow by the pipeline length on the sector to obtain the ratio. A low ratio corresponds to a sector with few consumers, such as countryside zones, a high ratio is the opposite. The program selects the best profiling using those criterion.

5. CONCLUSION

In this paper, we presented Scouter, a system that demonstrated its usefulness in the WAVES project for improving the contextualization of identified anomalies. The primary goal was to offer a generic system that can adapt to any Big Data platform whatever the engine used and without losing the capability to process various types of data sources. To achieve such target, we chose an approach based on data connectors relying heavily on messaging queue system and we offer multiple NLP functionalities such as topic extraction and sentiment analysis. Because of its easy-to-use Docker package, Scouter is already used in other prototypes at Atos and we are aiming to extend it with novel features such as ontology enrichment based on a dictionary of concepts and the identification of duplicate events coming from different data sources.

Finally, we plan to improve the implementation by supporting various ontology formats (e.g. ttl, N3, RDF/XML, etc.) and adding new data sources to fit most use cases (e.g. traffic information, etc.).

6. REFERENCES

- [1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 15:1–15:58, July 2009.
- [2] H. Khrouf, B. Belabbess, L. Bihanic, G. Képéklian, and O. Curé, "WAVES: Big Data platform for real-time RDF stream processing," in *3rd Stream Reasoning workshop co-located with 15th International Semantic Web Conference, Kobe, Japan.*, pp. 37–48, 2016.
- [3] S. de S Sirisuriya, "A comparative study on web scraping," *International Research Conference*, nov 2015.
- [4] J. Lovins, "Development of a stemming algorithm.," *Mechanical Translation and Computational Linguistics*, 1968.
- [5] S. Ellouze, M. Jaoua, and H. Belguith, "Machine learning approach to evaluate multilingual summaries," in *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, April 2017.
- [6] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. ACL workshop on Text Summarization Branches Out*, 2004.
- [7] D. J. O. H. A. Collomb, C. Costea and L. Brunie, "A study and comparison of sentiment analysis methods for reputation evaluation," tech. rep., Mar. 2014.
- [8] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit.," in *ACL (System Demonstrations)*, The Association for Computer Linguistics, 2014.

KNOWLEDGE RETRIEVAL STRATEGY FOR SATELLITES SYSTEM MONITORING BASED ON DATA ANALYTICS TECHNIQUES

C. Ciancarelli, A. Intelisano, S.G. Neglia

Thales Alenia Space

Via Saccomuro, 24 – 00131 Rome, Italy

carlo.ciancarelli@thalesaleniaspace.com

arturo.intelisano@thalesaleniaspace.com

ABSTRACT

The retrieval of knowledge from data with automatic procedures and algorithms is a challenging mathematical and data science process for several reasons, mainly because it requires deep insight in the datasets and, typically, the identification of the most suitable processing techniques and models fitting the data. In the field of satellite Systems operation, monitoring of spacecraft telemetry data is one of the most critical tasks to guarantee both the satellite lifetime and the required end-user applications, such as Earth Observation services. In this contest, present paper aims to propose an innovative approach to implement a tool able to analyze satellite telemetry data in order to infer anomalies probabilities, on a framework based on data analytics techniques, with the objective to make automatic provisions on possible occurrences of anomalies.

Index Terms—Bayesian networks, satellite telemetry data, knowledge retrieval, data analytics.

1. INTRODUCTION

The large amount of telemetric information collected during years of in-flight operational life allows the analysis of correlations among the different observables. The correlations and covariance analyses include as well the possibility to extend the causal connection inference among different subsystems, enlarging the perimeter of typical FDIR (Failure Detection Isolation & Recovery) and troubleshooting research. Computer aided techniques as well as data-mining algorithms (as Bayesian networks) allow to inspect covariance relations in a deeper way, allowing the automatic generation of Directed Acyclic Graphs (DAG) for the evaluation of probabilities evolutions with respect to measured “evidences” (i.e. the probabilistic parameters whose values have been perfectly assessed from the history). After the Bayesian networks (BN) definition, a key topic is the reduction of the necessary observables and of their intrinsic parameters, in order to reduce the computational time and the “over-fitting problem” of the related model. The advantage is in principle the possibility

to use such algorithms also “in-flight” for automatic FDIR and autonomous reconfiguration strategies.

The goal of the present paper is to analyze the predictive capability of such algorithms using known history for nominal and failure conditions, evidencing the sensitivity to the “anomalous” behaviors of parameters that, apparently, are inside the “green flags” conditions, as evaluated by control ground stations and on-board surveillances.

The proposed technique offers also the advantage of re-defining reliability estimations for the various subsystems analyzed “as a whole”, through their real interactions. The impact at design level can be a simplification of the design margins and of the redundancies strategies, going in the direction of an efficient “design to cost” approach (particularly useful for small satellites and big constellations).

Finally, the paper shows some simulation results related to the analysis of real in-flight telemetries and correlated operative modes.

2. BAYESIAN NETWORKS

A brief introduction on Bayesian networks [1]. BN utilize the probability calculus together with an underlying graphical structure to provide a theoretical framework for modeling uncertainty. Although the philosophical roots of BN may be traced back to Bayes and the foundations of probability [2], BN as such are a modern device, first appearing in Pearl (1988), and growing out of the research in expert or intelligent systems.

BN provide a comprehensive framework to model the dependencies between the variables in static data. BN are a class of *graphical models* that allow a concise representation of the probabilistic dependencies between a given set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ as a *Directed Acyclic Graph* (DAG) $G = (\mathbf{V}, A)$, where $\mathbf{V} = \{v_1, \dots, v_n\}$ is a set of nodes and A is a set of pairs of vertices called *arcs* (or *links*). Each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i , for $i = 1, \dots, n$. In other words, DAG defines a

factorization of a joint probability distribution over the variables that are represented by the nodes of the DAG, where the factorization is given by the directed links of the DAG.

Even though the joint probability distribution specified by a BN is defined in terms of conditional independence, BN are most often constructed using the notion of cause-effect relations. In practice, cause-effect relations between entities of a problem domain can be represented in a BN using a graph of nodes representing random variables and links representing cause-effect relations between the entities. Usually, the construction of a BN proceeds according to an iterative procedure where the set of nodes and their states and the set of links are updated iteratively as the model becomes more and more refined.

Moreover, BN are commonly considered as the knowledge base, i.e. a structured way to formulate our knowledge about the problem domain.

Above concepts have been applied in the paper to build up the proposed algorithms.

3. DATASET

The dataset has been set-up starting from the real satellite telemetry data generated by in-flight Earth Observation satellite system, that have been collected and stored during several years of its operational life time.

Satellite's telemetries provide a huge amount of information (typically, more than ten thousand parameters are defined). Although most of parameters can be immediately cut off from the data analysis, there would be still hundreds of parameters to be taken into account. Anyhow, processing and analyzing such data is too expensive and likely useless. Indeed, they are strongly affected by the presence of events (e.g. environmental phenomena or routine operations) which can bias all the dataset, increasing the correlation among parameters. Therefore, the a-priori knowledge on the satellite operations is mandatory to perform a proper selection of variables to build-up the dataset.

The satellite telemetry data that have been considered in the dataset concern the measurements of the following type of parameters:

- a) on-board equipment temperatures
- b) on-board equipment voltages and currents
- c) avionics parameters
- d) angular errors and angular rates
- e) reaction wheels speed.

An overall amount of around 30 variables (i.e. satellite telemetries) have been identified for the dataset, with the aim to obtain a comprehensive and significant set of parameters associated with the satellite attitude control system (platform avionics) and with the operative status of related on-board equipment units. The number of variables and the observation timeframe have obviously driven the

computational time needed to successfully process the dataset with the proposed algorithms. Therefore, the dimensions of the dataset have been assessed and tailored accordingly.

Actually the dataset has been assessed and refined also on the basis of the outcomes of the processing performed through the proposed BN algorithms. Indeed, the results of the first iterations have highlighted that some (few) variables showed a very limited correlation with the others and, therefore, have been considered of minor interest from knowledge retrieval point of view.

The following criteria have been applied to the whole dataset in order to allow the generation of suitable network topologies to be effectively used for the data analysis:

- a) Identification of relevant parameters. Only the variables which effectively enters in the sensitivity of the study of interest have to be included. Moreover, parameters that indicate temperatures in different positions of the spacecraft can be ignored.
- b) Among the relevant variables, select those providing clear information; for instance, it is preferred to use smoothed values for a given parameter (e.g. battery voltage) where the random noise is suitably filtered out.
- c) Include suitable calculated parameters. In case a parameter is not directly available (e.g. voltage of a given equipment), it can be computed from the voltage measurements available in telemetries.

As a consequence, the dataset has been updated, leaving basically unchanged the overall amount of variables. This methodology has provided the technical elements to refine it, i.e. to eliminate some variables and to include others that are recognized to be more correlated with observed events. As a result of this process, further steps of BN learning have been iterated, in order to obtain updated network topologies closer to the actual behavior of the satellite, i.e. better inferring the information flow and correlations among the selected variables. Topological modification of the Bayesian network structure has been profitably obtained with the support of technical specialists in satellite engineering.

As expected, this iterative approach in which preliminary results of BN structures are compared and assessed with the a-priori specific knowledge of the satellite operations, has provided a more refined BN model.

4. METHODOLOGY AND TOOLS

BN structure learning has been implemented by using the constraint-based algorithms [3], that learn the network structure by analyzing the probabilistic relations among variables and then constructing a graph. The action of these algorithms can be summarized in three steps:

1. learning of the network skeleton (the undirected graph underlying the network structure)

2. set all direction of the arcs that are part of a ν -structure (a triplet of nodes incident on a converging connection $X_j \rightarrow X_i \leftarrow X_k$)
3. set the directions of the other arcs as needed to satisfy the acyclicity constraint.

The processing of the dataset has been carried out with the following five constraint-based learning algorithms:

- a) *Grow-Shrink*: based on the *Grow-Shrink Markov blanket*, the simplest Markov blanket detection algorithm used in a structure learning algorithm;
- b) *Incremental Association*: based on the *Incremental Association Markov blanket* (IAMB) algorithm, which is based on a two-phase selection scheme (a forward selection followed by an attempt to remove false positives);
- c) *Fast Incremental Association* (Fast-IAMB): a variant of IAMB which uses speculative stepwise forward selection to reduce the number of conditional independence tests;
- d) *Interleaved Incremental Association* (Inter-IAMB): another variant of IAMB which uses forward stepwise selection to avoid false positives in the Markov blanket detection phase;
- e) *Max-Min Parents and Children* (MMPC): a forward selection technique for neighborhood detection based on the maximization of the minimum association measure observed with any subset of the nodes selected in the previous iterations. It learns the underlying structure of the BN (all the arcs are undirected, no attempt is made to detect their orientation).

Moreover, also the *Hill-Climbing* score-based algorithm has been applied, which assigns a score to each candidate BN and try to maximize it with heuristic search algorithm.

This methodology has been implemented through the software environment provided by R tool [4], equipped with the relevant packages implementing the BN model and associated learning algorithms, and applied to the satellite telemetry dataset.

5. BN SENSITIVITY ANALYSIS

The above described methodology has been applied to the dataset to generate the corresponding BN structures as input to the data analyst.

A first analysis has regarded the sensitivity of the BN network topology w.r.t. the dataset length (i.e. observation timeframe), maintaining fixed the number of variables; the objective has been to investigate the behavior of the BN structure when learned with different portions of the satellite telemetry dataset, especially the ones corresponding with high variability of given variables (e.g. angular rates) due to specific phenomena observed during satellite operational

lifetime. It has been observed that, with a given BN learning algorithm, the network graph tends to maintain the bulk structure, but anyhow features some changes regarding the position of the nodes and the arcs. The sensitivity of the network topology to dataset length and the elapsed time needed for algorithm convergence (see Table 1, BN learned with IAMB algorithm), appear to be mainly related to the dataset variability (along time direction) and to the correlation among different variables. High computational capacity is a must to investigate this type of behavior on huge dataset.

learning cases (dataset length increased by fixed step)	1	2	3	4	5
nodes	27	27	27	27	27
arcs	35	32	33	28	32
undirected arcs	2	4	10	10	3
directed arcs	33	28	23	18	29
average markov blanket size	3.56	3.11	3.48	2.89	3.04
average neighbourhood size	2.59	2.37	2.44	2.07	2.37
average branching factor	1.22	1.04	0.85	0.67	1.07
tests used in the learning procedure	4534	4625	4539	4493	4482
elapsed time (s)	21.24	49.74	45.21	29.61	94.71

Table 1: BN topology sensitivity example (IAMB algorithm)

The comparison among the selected learning algorithms has shown a strong sensitivity of the network topology w.r.t. the same dataset (see Table 2); it is clearly evidenced a difference in the capability to identify the number of directed versus undirected arcs in the possible reconstruction of a coherent DAG (Direct Acyclic Graph).

model	Grow-Shrink	IAMB	Fast-IAMB	Inter-IAMB	Max-Min Parent Children	Hill-Climbing
nodes	27	27	27	27	27	27
arcs	31	27	34	33	42	181
undirected arcs	10	11	1	6	42	0
directed arcs	21	16	33	27	0	181
average markov blanket size	3.18	2.67	3.48	3.93	3.11	20.74
average neighbourhood size	2.30	2	2.52	2.44	3.11	13.41
average branching factor	0.78	0.59	1.22	1	0	6.70
tests used in the learning procedure	2600	4406	1479	11552	1034	5837
conditional independence test	Pearson's Correlation	Pearson's Correlation	Pearson's Correlation	Pearson's Correlation	Pearson's Correlation	BiC

Table 2: BN learning algorithms comparison

The most interesting aspect is the capability, common basically to all the BN, to identify subnets related to the various “families” of telemetries (subnet of temperatures, of voltages, of attitude telemetries, etc.). The correlation among these subnets and so the “nesting” capability of the algorithm is also strong sensitive to the “time distance” among the analyzed dataset (maintaining fixed the number of samples); this aspect reflects the fact that constrained-based algorithms modify the topology estimation in function of the covariance estimation: when data are poorly correlated in time the single algorithm is able to identify the subnets but not to correctly link the subnets each other; on the other side, when the covariance increases thanks to the larger “time distance” among the samples, the algorithm starts to identify the links among the subnet, enhancing the BN prediction capability. In this context, Fast-IAMB algorithm has shown apparently a greater efficacy in generating directed-arcs minimizing the number of tests used in the learning procedure (see Figure 1).

The sensitivity of the topology to the various analyzed algorithms represents a problem for the data analyst, in the

identification of the best BN modeling. The “nesting” arcs help in identifying the skeleton to be used as guideline for the reconstruction of the most representative BN. For this purpose the followed approach has consisted in reconstructing an “ensemble” of various BN for different dataset, sparse in time and in samples dimension, in order to stimulate various orders of possible correlations among data.

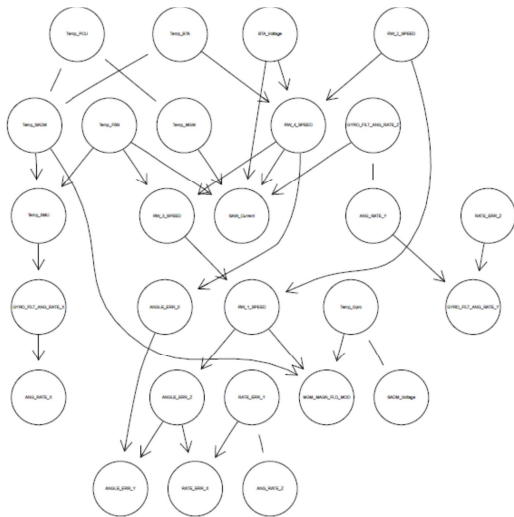


Figure 1: Fast-IAMB BN example of “nesting”

In fact the problem of predicting failures should require dynamic BN at the expense of an enormous computational effort. The approach described in this paper is based on a concept of “average BN”; at the end the different topologies reflect different covariances, different tests, different sensitivities to the datasets; the time becomes somehow “implicit” and frozen in the identified topology.

6. POSTERIOR ESTIMATE APPROACH

The set of Bayesian networks structures, obtained through the several learning algorithms and iterative processes described above, has been finally assessed to identify the best BN suitable to interpret the causal links among variables.

The identification of the “average BN” as a result of the trade-offs among algorithms and datasets described above, has been identified as the best approach to interpret the causal link among the variables. Each satellite telemetric datum has in fact a specific range of possible values; the average BN is then composed by a certain number of nodes identifying the “evidences”, i.e. the percentile of the values as result from the telemetries; some nodes are maintained as the “unknown”, being their estimated percentile the “result” from the Bayesian inference of the other nodes. Figure 2 provides an example of “average BN” for posterior estimation.

If a percentile, as evaluated by the network, is probabilistically out of the nominal range, than the result can be interpreted as the identification of a certain probability of failure. What is not possible to estimate in a precise way is the time to failure, because of the implicit use of time of the entire methodology.

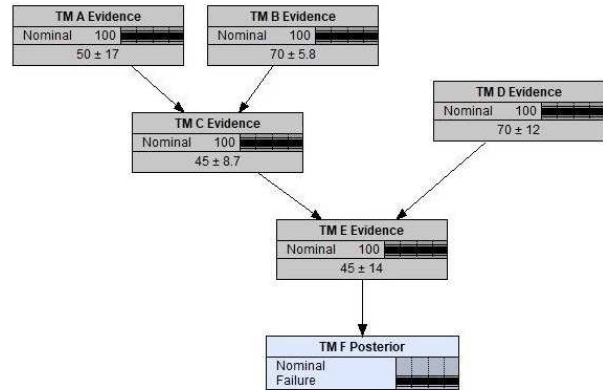


Figure 2: “Average BN” for posterior estimation

7. CONCLUSIONS

The paper focused on data analytics techniques based on Bayesian networks algorithms applied to real satellite telemetry dataset generated from in-flight Earth Observation satellite operational system. The proposed methodology provides the means to retrieve knowledge from the telemetry data, building up network topologies inferring the cause-effect relations among the dataset variables. A BN sensitivity analysis w.r.t. datasets and learning algorithms has been carried out. A method for the estimation of the posteriors probabilities has been finally identified as the best approach.

8. REFERENCES

[1] Uffe B. Kjærulff, Anders L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer, 2013.

[2] B. De Finetti, *Theory of Probability*, John Wiley & Sons, 1990.

[3] M. Scutari, J.B. Denis, *Bayesian Networks – With Examples in R*, CRC Press, NW, 2015.

[4] R Core Team (2017), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

ADAPTING EFFICIENTLY MID-LEVEL FEATURES TO HIGH RESOLUTION SATELLITE IMAGES INDEXING

Assia Kourgli¹, Lynda Bouchemakh¹, Aichouche Belhadj-Aissa¹ and Youcef Oukil²

¹USTHB, LTIR, Faculté d'Electronique et d'Informatique, Bab-Ezzouar, Alger, Algérie.

²ENS Bouzareah, Département de Géographie, Alger, Algérie

ABSTRACT

High-resolution satellite imagery (HRSI) permits to reach more accurate characterization of land cover, but in return, induces new challenges related to the transferring, archiving, analysis of this huge quantity of data. Recently deep features showed interesting performances compared to mid-levels features for HRSI retrieval task. Even if deep learning seems to be able to derive the most powerful features, it suffers from some drawbacks: unless transfer learning is used, it requires a large amount of training data. Moreover, the learning step is time consuming. In this paper, we show that by adapting mid-level features to HRSI nature and complexity, we obtain more representative features that permit to improved the HRSI retrieval features regard to precision and rapidity. The preliminary results obtained are better than those reached using key point descriptors namely SIFT and SURF features in terms of performances considering precision and time execution.

Index Terms— High-Resolution Satellite Images, Content-based Image Retrieval.

1. INTRODUCTION

One of the fundamental challenges with remote sensed big data such as HRSI is related to the retrieval aspect. Indeed, a huge amount of HRSI images are acquire and delivered every day. In this context, a content-based image retrieval (CBIR) scheme dedicated to HRSI must meet a number of criteria. The most important ones are expressed in terms of precision allied to rapidity. Over the last decades many approaches have been proposed, they differ mainly by the features used. These latter condition the retrieval performance. The more representative they are, the more the CBIR scheme is efficient. The features employed can be broadly divided in three classes: low-level features such as color, texture and texture [1-4] that are compared via a similarity measure [5], mid-level features obtained through the combination of rotation and scale invariant features such as SIFT (Scale Invariant Features Transform) and SURF (Speed-Up Robust Features) with codebook construction [6-8] and more recently, deep learning features derived from convolutional neural networks (CNNs). Even if deep

learning seems to be able to derive the most powerful features [9,10], it suffers from some drawbacks: unless transfer learning is used, it requires a large amount of training data. Moreover, the learning step is time consuming and particularly difficult with remote sensed images that contains very heterogeneous areas within the same category [4]. For these reasons, we still believe that mid-level features such as SIFT and SURF descriptors combined to bag-of-visual-words (BOVW) representation can challenge deep learning features specially when applied to the retrieval of HRSI. Indeed, the objects of interest may appear at different scales, orientation and illumination which leads to decrease the retrieval efficiency. Thus, to better deal with HRSI complexity, a new mid-level feature is introduced in this paper. It is derived from a modified HOG (Histogram of oriented gradients) representation [11] combined to BOVW quantification.

2. HOG-LIKE FEATURE

HOG feature permits to captures edge or gradient structure of local object appearance and shape. The classical HOG [11] decomposes an image into small squared cells, computes a local 1-D histogram of gradient directions in each cell, normalizes the result using a block-wise pattern for improved accuracy, and returns a descriptor for each cell. By concatenating all the normalized histograms into a single vector, the global HOG feature that is not invariant to rotation is obtained. To derive a feature adapted to HRSI, we considered a circular neighborhood (Figure 1) divided in 'N' parts.

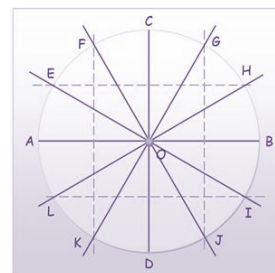


Figure 1. SC-HOG neighborhood structure

First, the global HOG feature is computed to determine the main direction θ and its corresponding part $n' \in [1, N]$. Then for each part of the neighborhood, a local HOG is computed according to the direction θ pre-computed. These local HOGs are reordered (circular shift beginning with n') and concatenated to form the shifted circular HOG (SC-HOG) that is invariant to rotation. Then color information is added by computing the histograms of RGB channels on the neighborhood defined above using b bins. The vector obtained is concatenated to the SC-HOG one to constitute SC-Color HOG (SC-CHOG).

Let us recall that mid-level features SIFT and SURF are computed for key local points determined by a multi-scale analysis [6]. Both descriptors produce different number of key points for each image according to some fixed thresholds making some classes less represented than others when aggregating the different descriptors via BOVW construction. To overcome this disadvantage and provide a more representative dictionary, the SC-HOG is computed on a dense grid using overlapping circular blocks. Then the different SC-HOG descriptors are clustered to produce the codebook. This quantification step is reached through the use of k-means algorithm. The centroids obtained constitute the visual words. Then each image is represented by its histogram of visual words.

3. RESULTS

The dataset UCM used is a manually constructed one [6] consisting of 21 image classes LULC (Land use Land cover), containing each 100 images of size 256×256 (Figure 2) with spatial resolution of 30 cm [6]. It contains the following classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court.

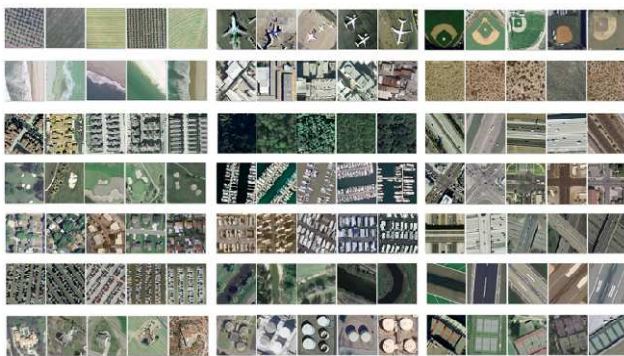


Figure 2. Samples of UCM dataset [6]

The descriptors CS-HOGs are computed over the different images and then quantified via a BOVW to create a 150 bin histogram. While for CS-CHOG, the histogram of the three colors channels using 4 bins is added to the global HOG.

In a query by example scheme, we are interested in retrieving several similar images by comparing two descriptors (in this context the 150 bins histogram) to obtain a measure of similarity through a distance measure. The metrics tested are reported in table 1.

Table 1. Common Similarity measures

Manhattan distance	$(L1) = d(x, y) = \sum_{i=1}^n x_i - y_i $
Euclidian distance	$(L2) = d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
Chi-square distance	$d(x, y) = \sqrt{\sum_{i=1}^n \frac{(x_i - y_i)^2}{(x_i + y_i)}}$
Canberra distance	$d(x, y) = \sum_{i=1}^n \frac{ x_i - y_i }{ x_i + y_i }$
Squared chords distance	$d(x, y) = \sum_{i=1}^n (\sqrt{x_i} - \sqrt{y_i})^2$

At this step, a texture information is added. Accordingly, a weight is incorporated to the similarity measure :

$$d(x, y) = \alpha \cdot d(x, y)$$

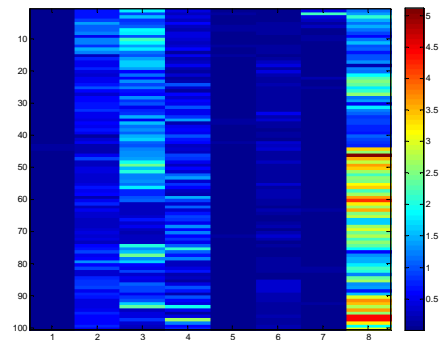
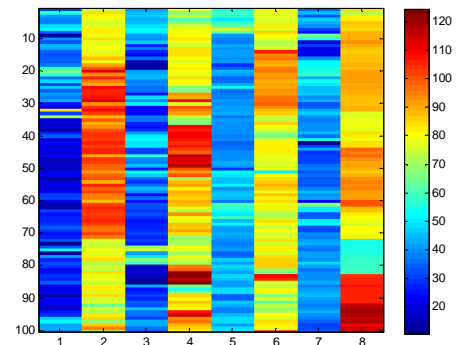


Figure 3. Mean Range of local variance (up) and variance of local entropy (down) for each sample (1: agricultural, 2: airplane, 3: beach, 4: buildings, 5: chaparral, 6: dense residential, 7: forest, 8: harbor).

α is simply obtained through the comparison of a texture measure such as variance and entropy (See Fig.3). For each query image the distance is weighted by the sum of the difference of the texture parameters.

Usually, CBIR performance is measured by precision and recall, as well as precision-recall curves [4].

$$P = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}}$$

$$AP = \frac{1}{Nq} \sum_{i=1}^{Nq} P(i)$$

where Nq represents the number of queries.

Similarly, recall R and average recall AR are given as :

$$P = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images}}$$

$$AR = \frac{1}{Nq} \sum_{i=1}^{Nq} R(i)$$

To show the performance of the descriptor proposed, we give some preliminary results. Thus, tests have been conducted using 8 classes according to the following reference [3]. These are 1: agricultural, 2: airplane, 3: beach, 4: buildings, 5: chaparral, 6: dense residential, 7: forest, and 8: harbor.

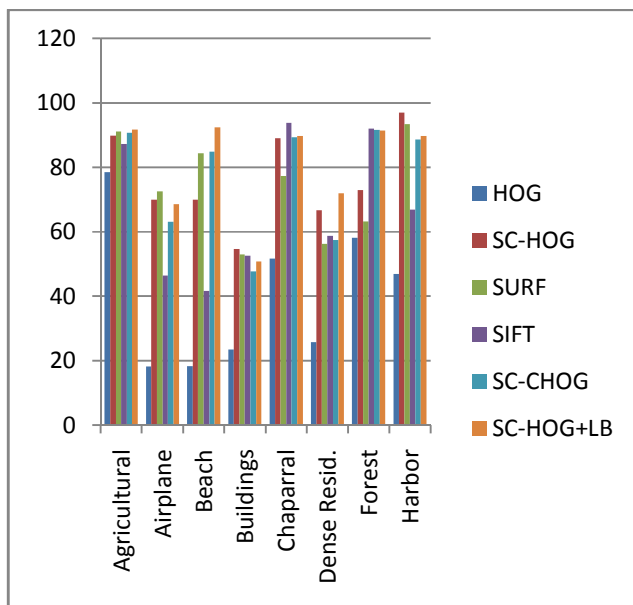


Figure 4. Average precisions for 8 classes

Figure 4 gives the average precision for 8 categories for the original HOG, SIFT, SURF and SC-HOG, SC-CHOG where

SC-HOG+ LB corresponds to the case where texture is added to the similarity measure acting as a labeling [12]. All the descriptors are aggregated with the same number of clusters for comparison purposes. In most cases, the SC-CHOG leads to the highest average precision values. Its performance is boosted when texture is incorporated (SC-CHOG+LB) in the labeling process.

Figure 5 illustrates precision-recall curves obtained using the 800 images. These curves show that SC-HOG is the most efficient. Instead of giving some visual examples, we present three maps (Figure 6) each one summarize all the retrieval results coded in different color. Each line should be read from left to right. For example, the 100 first lines correspond to the retrieval results for agricultural class. One can see that code is blue for the first images retrieved (beginning by the left) which means that the images belonging to agricultural classes are well retrieved. What is interested to note here is that when using SC-CHOG descriptor the following 100 images (from 100 to 200) are coded in orange corresponding to forest category. Thus the next images retrieved correspond to forest that is the most similar class to agricultural one. The same can be observed for chaparral and forest classes.

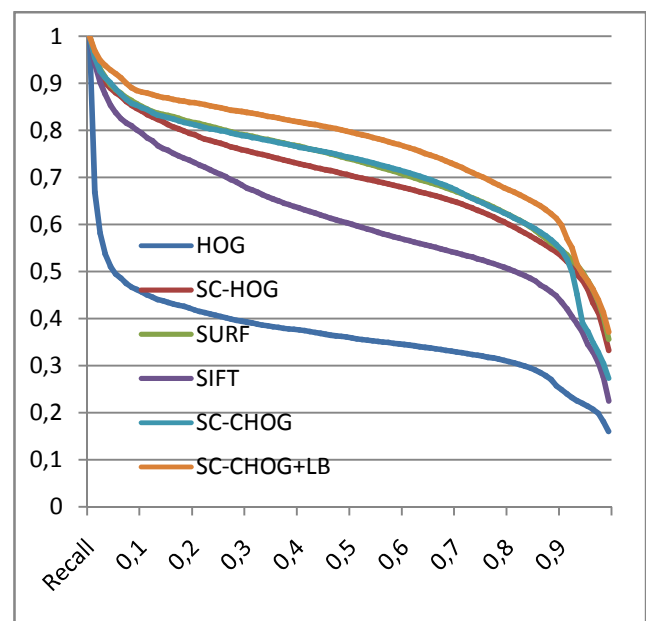
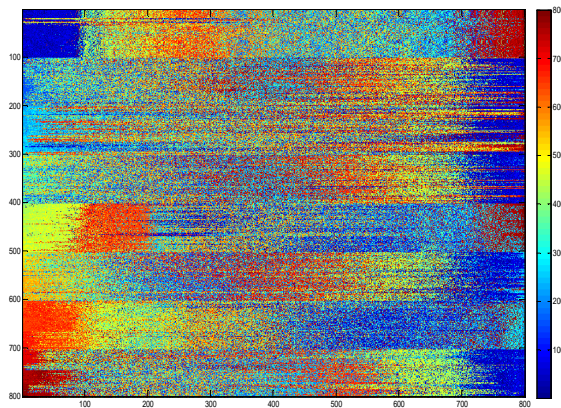


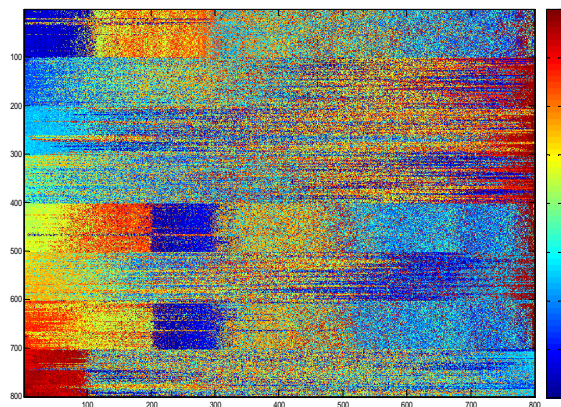
Figure 5. Precision-recall curves for 8 classes

One can also observe that beach and buildings categories (coded in light blue) are better retrieved when using the descriptor proposed.

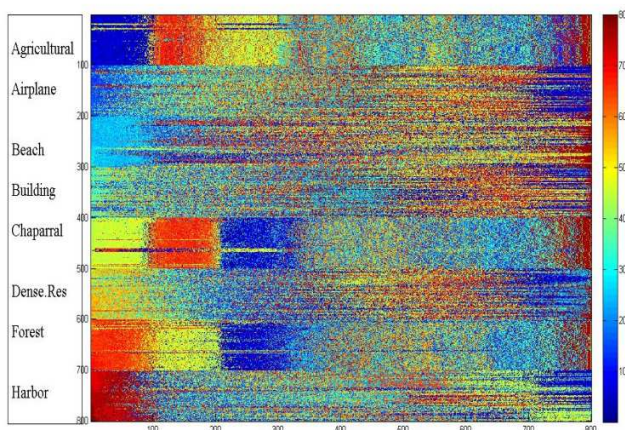
These preliminary results to show that SC-HOG descriptor performs better than classical descriptors based on key points detection that are more time consuming.



a) Retrieval results using SIFT+LB descriptor



b) Retrieval results using SURF+LB descriptor



c) Retrieval results using SC-CHOG+LB descriptor

Figure 6. Retrieval results maps

4. CONCLUSION

The recent successes of deep learning approaches have occulted the progresses that are still possible to reach when using mid level features in retrieval schemes. In this paper, we propose to compute invariant to rotation features that are derived from HOG representation. To deal with HRSI complexity, we added color and texture information to the retrieval scheme. We quantitatively analyzed the efficiency of this scheme that yields to better result than usual mid level features namely SIFT and SURF. Let us note that the SC-HOG descriptor is a more compact and thus constitute a representative feature that permits to improved the HRSI retrieval features regard to precision and rapidity. Results obtained here are only preliminary. Further tests must be conducted. Future work includes extending the investigation to consider the 21 categories to generalize the process..

5. REFERENCES

- [1] M. Datcu, K. Seidel, and M. Walessa, "Spatial information retrieval from remote-sensing images. i.information theoretical perspective", *IEEE Trans. on Geosciences and Remote Sensing*, 36(5), pp. 1431-1445, 1998.
- [2] S. Aksoy, S., I.Z. Yalniz, J. Nick, K. Tasdemir. "Automatic detection and segmentation of orchards using very high-resolution imagery", *IEEE Trans. on Geosciences and Remote Sensing* 50, pp. 3117–3131, 2012.
- [3] Z. Shao, W. Zhou, L. Zhang, J. Hou, "Improved colour texture descriptors for remote sensing image retrieval", *Journal of Applied Remote Sensing*, 8(1), 2014.
- [4] H. Sebai, A. Kourgli, "Dual-tree complex wavelet transform applied on colour descriptors for remote-sensed images retrieval", *Journal of Applied Remote Sensing*, 9(1), 2015.
- [5] Q. Bao, P. Guo "Comparative studies on similarity measures for remote sensing image retrieval", In: IEEE Inter. Conference on Systems, Man and Cybernetics, pp. 1112–1116, 2004.
- [6] Y. Yang, S. Newman, "Geographic image retrieval using local invariant features", *IEEE Trans. Geosciences and Remote Sensing*, pp. 818–832, 2012.
- [7] E. Eptoula, "Remote sensing image retrieval with global morphological texture descriptors", *IEEE Trans. on Geosciences and Remote Sensing*, 52(5), 2014.
- [8] S.Ozkan, T. Ates, E. Tola, M. Soysal, and E. Esen, "Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization", *IEEE Geosci. Remote Sensing Letters* 11(11), pp. 1996-2000, 2014.
- [9] P. Napoletano, "Visual descriptors for content-based retrieval of remote sensing images", CoRR, vol. abs/1602.00970, 2016.
- [10] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang. AID: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Trans. on Geoscience and Remote Sensing*, Vol.55, No.7, pp. 3965 - 3981, 2017.
- [11] N. Dalal, B.Triggs, "Histograms of oriented gradients for human detection", In: Proceedings of the IEEE Computer Society Conference on Comp. Vis. and Patt. Recog.1, pp. 886–893, 2005.
- [12] A. Kourgli, H. Sebai, S. Bouteldja, Y.Oukil: "Towards adaptive high-resolution images retrieval schemes". *The International Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume XLI-B2, 2016. p.201-209, DOI: 10.5194/isprs-archives-XLI-B2-201-2016

TOWARDS A MAP OF THE EUROPEAN TREE COVER BASED ON SENTINEL-2

Thor-Bjørn Ottosen¹, Geoffrey Petch¹, Mary Hanson¹ and Carsten Ambelas Skjøth¹

¹Institute of Science and the Environment, University of Worcester, Worcester, UK

ABSTRACT

Many areas of science and policy depend on knowledge of the tree cover in Europe. Sentinel-2 is a new (launched in 2015) satellite with a higher spatial resolution compared to previous satellites. In the present study a new algorithm for mapping tree cover from Sentinel-2 is developed, an analysis of which bands should be used for tree cover mapping is made, the accuracy of the mapping is assessed, and the tree cover from the present approach is compared with previous estimates. Firstly, the feasibility of the present algorithm is demonstrated. Secondly, it is shown that only ten band combinations have good performance in four selected Sentinel-2 tiles and that the bands 3, 5, 6, 12 appear in most combinations. Thirdly, the accuracy is assessed to be high, and lastly it is shown that the relative difference between the tree cover of the present study and the tree cover of previous studies is between -14% and 68%.

Index Terms— *Algorithm design, spectral bands, broadleaved trees, coniferous trees, Intel DAAL.*

1. INTRODUCTION

The European tree cover is important for many areas of science and planning such as climate [1], atmospheric composition [2, 3] and socio-economic values [4]. Until recently, the best estimate of the European tree cover has been $1.47 \cdot 10^6 km^2 \pm 0.02 \cdot 10^6 km^2$ based on either a combination of MODIS images and LISS-3 images [5] or Landsat images [6] with a spatial resolution of 25 m – 30 m. In old cultural landscapes, like the UK and other European countries, trees are often located in small linear features or groups and along roads and rivers, thus making them difficult to detect using satellites such as Landsat. It is thus likely that the actual tree cover is higher than previously thought and that woodland is more widely distributed.

With the launch of Sentinel-2, a new tool has become available with a high spatial resolution and a band combination specifically designed for vegetation studies. However, a number of methodological questions remain to be answered before a European tree map can be produced from Sentinel-2. In this paper we explore a methodology for tree cover mapping from Sentinel-2, we analyse the performance of all 8100 potential band combinations for tree mapping,

assess the accuracy of the methodology and compare with previous estimates for four Sentinel-2 tiles.

1. METHODS

1.1. Mapping tree cover

The tree mapping algorithm consists of a number of steps carried out sequentially:

1. Clouds and other artifacts are removed using the accompanying masks.
2. All bands are resampled to $10 m \times 10 m$.
3. All bands are normalized by mean centering and division with the standard deviation. This is done to normalize the weight of the individual bands.
4. A K-means unsupervised classification is done using Intel Data Analytics Abstraction Library (DAAL) with the number of classes set to 25.
5. The NDVI values of the pixels corresponding to forests in Corine Land Cover (CLC) [7] are extracted from the image.
6. Pixels from the entire image are subsequently removed if their NDVI value is less than $\mu - 2\sigma$ of the distribution of values extracted in step 5. In this way, non-vegetation pixels are efficiently removed.
7. The dominating classes for respectively broadleaved and coniferous forests are labelled using CLC by K-means clustering the clusters from the first clustering into two classes (dominating and non-dominating). This is done iteratively starting from the largest polygons within the Sentinel-2 tile proceeding to the smallest polygon until convergence. Convergence is defined when the largest change in a class is smaller than 1%.

1.2. Analysis of band combinations

To analyze whether all spectral bands were needed for this algorithm, and if not, which spectral bands provided the best performance, the algorithm was applied to all 8100 possible band combinations of size 3 to 13 on four Sentinel-2 tiles covering selected areas of Northern Europe. These tiles are named 30UWC, 30VUH, 32NVH and 33VUC. The kappa-coefficient for the wall-to-wall comparison with the respective national forest inventory was subsequently calculated and the performance of each band combination

ranked by summing the kappa coefficient over the four images.

1.3. Accuracy assessment

The accuracy assessment of the present study was performed on tile 30UWC.

1.3.1. Sampling design

1000 pixels were selected across the image using stratified random sampling [8]. The pixels were stratified into *broadleaved trees*, *coniferous trees* and *no trees* to prevent the non-forest category dominating the results.

1.3.2. Response design

The primary land use class of each pixel was subsequently manually determined using Google Earth. The interpreter did not have access to the classified map during the manual classification to avoid biasing the classification (blind interpretation). To enhance consistency among interpreters, a written guide to the classification procedure was produced and 99 points were classified by all interpreters. As well as *broadleaved trees*, *coniferous trees* and *non-forest*, the interpreter had opportunity to classify a pixel as *unclassified* and *unclassified trees*. Data points that were classified to be in the last two categories were subsequently excluded from the analysis.

2. PRELIMINARY RESULTS AND DISCUSSION

2.1. Mapping tree cover

Figure 1 shows an example of a result of the labelling procedure of the 25 classes produced by the unsupervised classification algorithm. The blue bars are the areas marked as broadleaved forest in CLC and the orange bars are the areas marked as coniferous forest in CLC. In this case, classes 18 and 22 dominate broadleaved forests, whereas classes 8 and 12 dominate coniferous forests. It is likewise evident, that these four classes are the dominant ones and therefore the ones labelled as forests since a complete separation cannot be achieved as the CLC land cover by definition will miss small woodland areas outside forests and miss small bare patches inside forest regions.

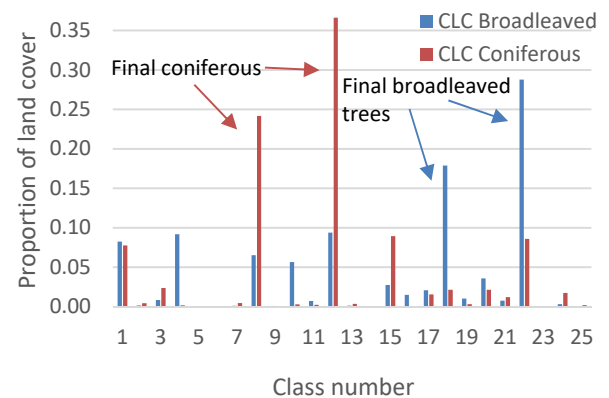


Figure 1 Example of a result of the labelling procedure.

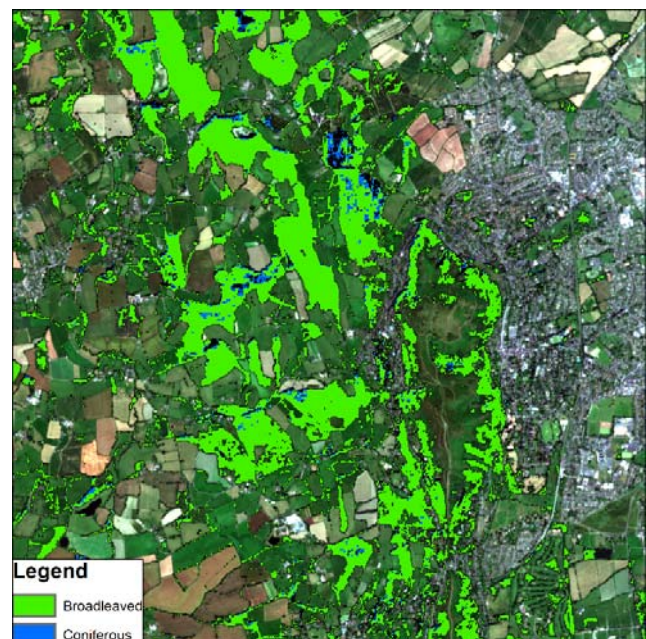


Figure 2 Example of a tree map on the border between Herefordshire and Worcestershire, UK (tile 30UWC) superimposed on a RGB-image from Sentinel-2. The scale is 1:75,000.

An example of a classification result is shown in Figure 2 superimposed on a RGB-image from Sentinel-2. The area is dominated by broadleaved forests with only small patches of coniferous forests. It is evident from this figure, that the algorithm allows the mapping of trees with a high degree of detail. This is seen from the many small features in the left side of the image and the trees mapped in the urban area in the right side of the image. These small features are by definition not a part of the CLC land cover or the more detailed UK tree cover map. Furthermore the map also contains a separation into coniferous and broadleaved trees.

2.2. Analysis of band combinations

Table 1 shows the ten band combinations that appear among the top-5% combinations in all four images. It is evident that bands 2, 3, 6 and 12 appear in many of the combinations. Band 2 is the blue band (496.6 nm), band 3 is the green band (560.0 nm), band 6 is a red-edge band (740.2 nm), and band 12 is a short-wave infrared band (2202.4 nm). Using USGS Spectral Characteristics Viewer it can be seen that these bands are particularly suitable to separate different types of vegetation.

Table 1 Band combinations appearing in top-1500 in all four images and their corresponding kappa-sum.

$\sum \kappa_i$	Combination
2.755	1, 2, 3, 4, 5, 7, 9, 12
2.803	2, 3, 6, 12
2.774	1, 3, 5, 6, 12
2.771	1, 2, 3, 5, 6, 12
2.800	1, 3, 4, 5, 6, 11, 12
2.797	2, 4, 5, 6, 12
2.799	3, 5, 6, 11, 12
2.797	2, 3, 4, 5, 6, 11
2.757	3, 4, 5, 6, 7, 9, 11, 12

Out of these combinations, the combination 2, 3, 6 and 12 was chosen, since it has a good performance in all four images.

2.3. Accuracy assessment

Table 2 Confusion matrix for the satellite map for tile 30UWC versus google earth. 0 is non-forest, 1 is broadleaved forest, 2 is coniferous forest, n=941.

Reference data	Classified data			Producer's Acc. (%)
	0	1	2	
0	303	67	18	0.78
1	23	228	182	0.53
2	3	10	107	0.89
User's Acc. (%)	0.92	0.75	0.35	0.68

The results of the accuracy assessment are shown in Table 2. Compared to the commonly used success criteria of 85% accuracy, an overall accuracy of 68% is low. The algorithm is especially challenged in separating broadleaved trees from coniferous trees. This is natural, since these two classes are spectrally much more similar compared to the non-forest class. This shows that the map tends to overestimate the cover of coniferous forest. Part of this is the result of shadows on forest edges, forest roads etc. that gives the trees a darker color. Future work should aim at reducing this effect. A group of points are classified as forests, and referenced as non-forest. This is largely caused by green fields being spectrally

similar to forests. In the future we will implement a temporal averaging procedure over several images, which we hope will reduce this effect. The accuracy for the map (area \times accuracy) is 0.76 or 0.90 depending on whether the user's accuracy or the producer's accuracy is used. Averaging the two numbers gives 83%, which is very close to the commonly used limit. The map accuracy is 88% if classified as only *trees* and *no trees*. Considering that the accuracy assessment is done on 10 m \times 10 m resolution, which is higher than previous maps, this result must be considered acceptable.

2.4. Comparison with previous datasets

Table 3 compares the amount of tree cover in the dataset of [5] with the tree cover mapped from Sentinel-2 for four tiles in Northern Europe. As can be seen, the relative difference is between -14% and 68% with Sentinel-2 often estimating a larger tree cover. Generally, this looks like a realistic result. Widespread clouds, where the accompanying cloud mask does not remove some of them, influence 32VNH. It is thus plausible that this value is an overestimation. That Sentinel-2 has the lower of the two numbers in 33VUC is likewise an effect of high clouds that have not been removed by the cloud mask. This demonstrates the ability of Sentinel-2 to detect smaller groups of trees compared to previous satellites. It is expected that using several images over the same area will either completely remove or at least reduce the effect of clouds, thus obtaining more accurate estimates of the tree cover in cultural landscapes over very large areas like Europe.

Table 3 Comparison between the tree cover in our map and the tree cover in the map of [5].

Tile code:	Location:	Our map (km ²)	Previous maps (km ²)
30UWC	Worcester, UK	1148	909
30VUH	Scotland	2204	1407
32VNH	West Denmark	2387	1178
33VUC	East Denmark	2175	2515

3. CONCLUSION AND OUTLOOK

The present study developed an algorithm for automatic tree cover mapping based on Sentinel-2 and demonstrated that this is a feasible approach to tree cover mapping, analyzed the performance of 8100 band combinations, assessed the accuracy, and demonstrated that Sentinel-2 can contribute to more accurate tree cover maps of Europe, in particular by identifying smaller woodlands that are important in some areas. In the future, this algorithm will be extrapolated to the entire European domain, to obtain a complete estimate of the European tree cover divided into broadleaved and coniferous trees.

4. ACKNOWLEDGEMENTS

The present study has been supported by the BBSRC funded project *New approaches for the early detection of tree health pests and pathogens*, projectID: BB/L012286/1.

5. REFERENCES

- [1] Bonan, G.B., “Forests and Climate Change: Forcings, Feedbacks, and the Climate Benefits of Forests.” *Science* (80-.). 320, 1444 LP-1449, 2008.
- [2] Kesselmeier, J., Staudt, M., “Biogenic Volatile Organic Compounds (VOC): An Overview on Emission, Physiology and Ecology.” *J. Atmos. Chem.* 33, pp. 23–88, 1999.
- [3] Pauling, A., Rotach, M.W., Gehrig, R., Clot, B., “A method to derive vegetation distribution maps for pollen dispersion models using birch as an example”. *Int J Biometeorol* 56, pp. 949–958, 2012.
- [4] FAO, *Global Forest Resources Assessment 2015*, FAO Forestry, 2015.
- [5] Kempeneers, P., Sedano, F., Seebach, L., Strobl, P., San-Miguel-Ayanz, J., “Data fusion of different spatial resolution remote sensing images applied to forest-type mapping.” *IEEE Trans. Geosci. Remote Sens.* 49, pp. 4977–4986, 2011.
- [6] Hansen, M.C., Potapov, P. V, Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S. V, Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., “High-Resolution Global Maps of 21st-Century Forest Cover Change.” *Science* (80-.). 342, 850 LP-853, 2013.
- [7] EEA, *CORINE land cover technical guide – Addendum 2000*, EEA, Copenhagen, 2000.
- [8] Stehman, S. V., “Sampling designs for accuracy assessment of land cover”. *Int. J. Remote Sens.* 30, 5243–5272. 2009.

BIGEARTH-ACCURATE AND SCALABLE PROCESSING OF BIG DATA IN EARTH OBSERVATION

Begüm Demir

Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
e-mail: demir@disi.unitn.it

ABSTRACT

This paper presents the BigEarth that is a research project funded by the European Research Council (ERC) Starting Grant for the period of 2018-2023. The BigEarth project aims to address very important scientific and practical problems by focusing on the main challenges of big earth observation data on remote sensing (RS) image characterization, indexing and search from massive archives. In particular, we develop novel methods and tools, aiming to: 1) characterize and exploit high level semantic content and spectral information present in RS images; 2) extract features directly from the compressed RS images; 3) achieve accurate and scalable RS image indexing and retrieval; and 4) integrate feature representations of different RS image sources into a unified form of feature representation. Moreover, a benchmark archive with high amount of multi-source RS images will be constructed. BigEarth is an interdisciplinary project between image processing, machine learning and remote sensing. Outputs of BigEarth will highly contribute to the climate change and ecological studies.

Index Terms—content based image retrieval, scalable image search, indexing, big data, remote sensing

1. INTRODUCTION

During the last decade, more and more satellites with optical and Synthetic Aperture Radar sensors onboard have been launched, and the developments of satellite technology has increased the variety, amount, and spatial/spectral resolution of Earth Observation (EO) data. Accordingly, huge amount of remote sensing (RS) images have been acquired, leading to massive EO data archives from which mining and retrieving useful information are challenging. In view of that, content based image retrieval (CBIR) has attracted great attention in the RS community [1]. The applications of querying image contents from large EO data archives and indexing them rely on the capability and effectiveness of: 1) the feature extraction techniques in describing RS images and their specific properties on spatial and spectral resolutions of the data; and 2) the retrieval algorithms in evaluating the similarity among the considered features [2]. The high spatial resolution images acquired by the new generation of satellite sensors (e.g., Sentinel 1 and 2 images) and high spectral

resolution images (e.g., hyperspectral images) require robust feature extractors that exploit the high information content. These feature extractors need to be designed by taking into account also the computational complexity to process large EO data. RS image feature extraction methods in literature have several limitations. All the methods either i) extract the image features on each single image band separately and then concatenate them to define a final image descriptor, or ii) extract the features only from a single image band that is defined by some feature extraction methods (such as Principle Component Analysis). Considering that each image band can be also modeled by multi-features, a simple concatenation of features of different image bands leads to highly increased feature dimensionality and thus a significant increase in the computational complexity of the retrieval phase. Moreover, some feature representations may be highly correlated and thus redundant. Due to the dramatically increased volume of RS image archives, it is required to compress the RS images before storing them in any storage devices. To characterize an image for retrieval problems, existing RS CBIR methods and tools require decoding images before applying any feature extraction method. Thus, they cannot be directly applied to compressed streams of EO data. However, this process is impractical due to highly increased computational time requirements in the case of large-scale RS image search and retrieval problems. Moreover, most of the existing RS CBIR methods and tools exhaustively compare the query image with each image in the archive (linear scan), which is time demanding and not scalable in operational applications [1]. Thus, the development of fast and accurate RS CBIR methods and tools is highly needed. Furthermore, in large-scale CBIR, the storage of the image descriptors in the auxiliary archive is also challenging as RS image contents are often represented in very high dimensional features. The availability of increased number of multi-source/multi-modal images (multispectral, hyperspectral and SAR) associated to the same geographical area motivates the need for effective methods, which can combine/fuse the image descriptors to define rich characterization of RS images (and thus to improve image retrieval performance). However, in RS CBIR scientific literature and existing tools, their fusion for RS CBIR problems has not been explored, yet. In addition, one of the critical issues in the framework of RS CBIR is to reach suitable image archives with reliable reference samples (i.e., annotated images) for the validation/test of the software and algorithms. There are only few benchmark archives, which

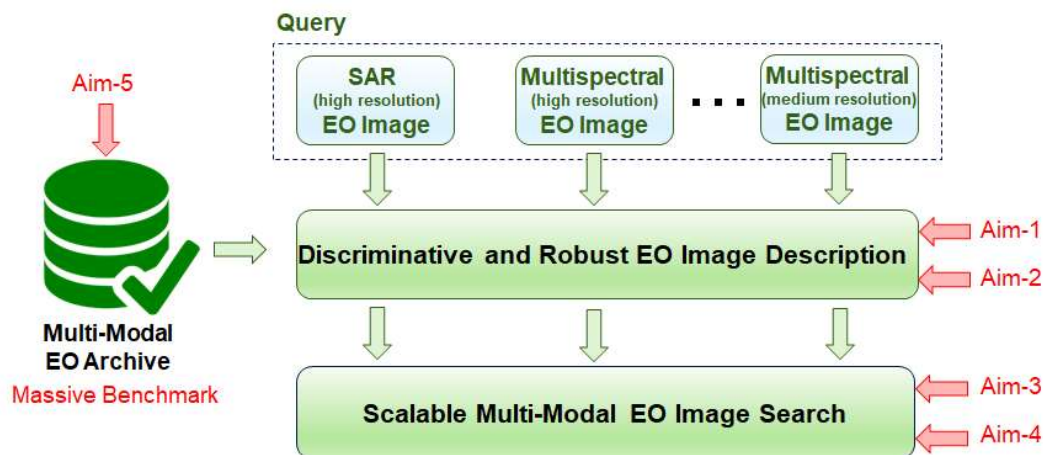


Fig . 1: BigEarth novel vision

include very small number of annotated images. Analysis of ‘Big Data in Earth Observation’ is a challenge that requires the development of innovative data processing methods for rapid and accurate content-based access to existing EO archives.

2. AIMS OF THE BIGEARTH PROJECT

Main objective of the BigEarth project is to develop highly innovative feature extraction and content based retrieval methods and tools for RS images, which can significantly improve the state-of-the-art both in the theory and in the tools currently available for RS CBIR problems (see Fig. 1 for the novel vision of the BigEarth). To this end, challenging and very important scientific and practical problems will be addressed by focusing on the main challenges of Big EO data, which are: RS image characterization, indexing and search from massive archives.

The BigEarth project consists of five Aims in total, from which four Aims are associated to the development of novel methodologies and tools on the main challenges of Big EO data and also one Aim is related to the benchmark archive construction to validate the algorithms and the software.

Aim 1: Development of novel methods and tools to characterize and exploit high level semantic and spectral information present in RS images;

Aim 2: Development of novel feature extraction methods and tools to directly extract features from the compressed RS images;

Aim 3: Development of accurate and scalable RS image indexing and retrieval methods together with associated tools;

Aim 4: Development of methods and tools to integrate feature representations of different RS image sources into a unified form of feature representation;

Aim 5: Construction of a benchmark archive with high number of multi-source RS images.

Schematic representation of the Aims of the BigEarth project is given in Fig. 2 and methodologies being developed for each Aim are explained in the following.

2.1. Aim 1: Differently from the feature extraction methods available in the scientific literature and existing tools in RS CBIR problems, novel feature extraction methods will be developed to: i) precisely characterize the detailed level of semantic information present in RS images; and ii) accurately describe the high spectral information of RS images for their characterization. To characterize high level of semantic information, the developed methods will be devoted to: 1) automatically categorize each image in the archive into several land-cover classes (i.e., primitive classes) by assigning multi-labels; 2) properly consider the co-occurrence of the multi-labels through modeling the label relations as a prior knowledge; and 3) retrieve images very similar to the query image by modeling effectively the high-level semantic content of RS images. To this end, a particular attention will be given in developing semi-supervised graph-theoretic methods in the context of multi-label RS image retrieval. These methods will include: i) RS image segmentation and feature extraction; ii) multi-label image categorization based on neighborhood graphs; and iii) image retrieval based on graphs. The methods being developed will require that a very small fraction of images in the archive is initially annotated as training images with their primitive class labels. A particular attention will be devoted to develop effective region feature extraction methods that: 1) can be applied to large amounts of EO data with low-computational complexity and can be useful on various kinds of EO data (i.e., optical and SAR); and 2) can accurately model high-spectral/spatial information in a given region. Development

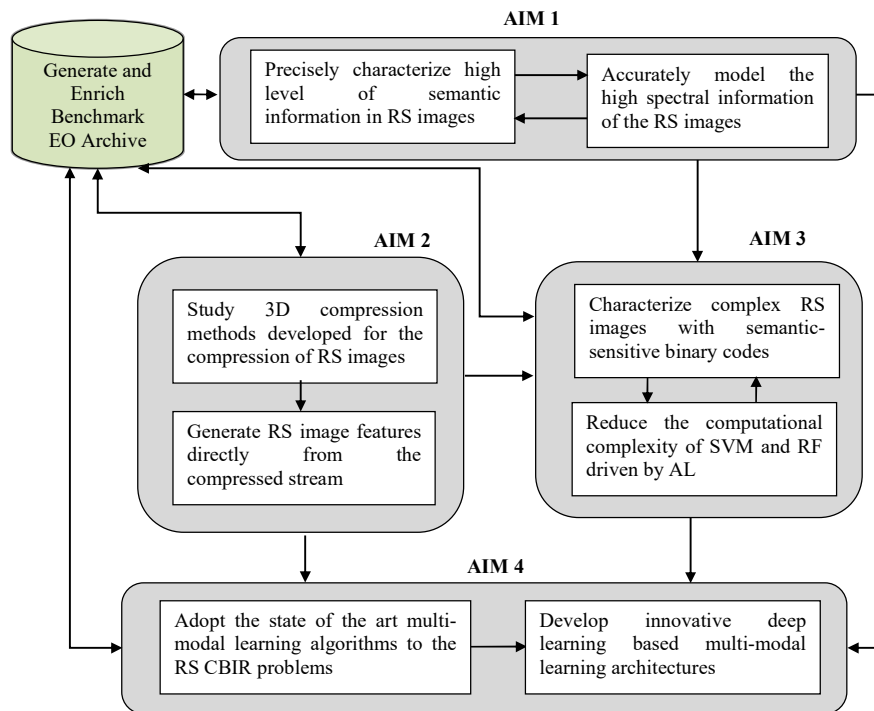


Fig. 2: Schematic representation of the Aims of the BigEarth Project

of fast parallel algorithms in the context of multi-label categorization and approximate graph matching to speed up the process will be considered. It may be possible that some primitive class labels can be absent in the training set. In order to overcome this problem, a scalable AL method through RF will be introduced in the context of the developed graph based methods. In order to accurately model the high spectral information of the RS images, innovative methods will be developed. On the one hand, the adaptation of well-known local invariant feature extraction methods (i.e., SIFT [3], HOG [4]) to a very high dimensional EO data will be investigated. On the other hand, novel methods will be developed aiming to characterize a RS image with a spectral descriptor (which will summarize the information existing in the spectral signatures of the image) and its combination with a spatial descriptor (which will summarize the spatial information of the image) will be investigated.

2.2. Aim 2: Due to the dramatically increased volume of RS image archives, it is required to compress the images before storing them in any storage devices. There are several 3D compression algorithms developed for the compression of multispectral and hyperspectral RS images in the literature [5-7]. To characterize an image for retrieval problems, existing RS CBIR methods/tools require decoding images before extracting features. However, this process is impractical due to highly increased computational time in large-scale RS CBIR problems. In order to reduce the computational time, methods/tools that are capable of

generating RS image features directly from the compressed stream without (or partially) performing the decoding process will be developed during the project. In this project, 3D RS image compression algorithms (which consider jointly and also separately spectral and spatial redundancies) existing in the RS literature will be initially considered. Then, promising feature extraction methods that can accurately characterize the encoded RS images by these methods will be developed. A particular attention will be given to multidimensional wavelet-based compression algorithms and investigation of feature extraction methods from wavelet coefficients. It is worth emphasizing that in RS there is no tool and no methodology introduced for CBIR problems that can extract features from 3D compressed domains.

2.3. Aim 3: To provide accurate and scalable RS image indexing and retrieval systems, innovative hashing-based methods and related tools will be developed to: i) characterize complex RS images with compact binary codes; and 2) reduce the computational complexity of RF driven by AL and the supervised classifier. The Aim 3 will take as input the outcomes (i.e., feature extraction methods) of the Aim 1-2. For this aim, novel semantic-sensitive hashing methods will be developed, which can effectively characterize high-level semantic concepts present in RS images. These methods will aim to represent a RS image by multiple hash codes, each of which corresponds to a primitive class. In details, different semantic-sensitive hashing methods that are: i) unsupervised; ii) semi-supervised; and iii) supervised will be developed that

can learn primitive-class-aware hash codes for each image. A particular attention will be given to develop methods that: 1) can be used with any kind of RS images (e.g. optical, SAR images); and 2) are effective to define accurate semantic-sensitive hash functions particularly for high-dimensional RS image descriptors. In RS, methods/tools for the primitive-class-aware hashing do not exist. To reduce the computational complexity of the RF driven by AL schemes, scalable AL methods will be developed based on the adaptation of hashing methods for this purpose. In addition to the hashing-based scalable AL, the tools with additive kernel SVMs (AK-SVMs) will be also developed with their exact and approximate solutions to speed up the retrieval process. Note that in recent years the AK-SVMs have gained an increasing interest for image retrieval problems in the computer-vision communities, whereas there is no tool available with AK-SVMs for RS CBIR problems and also scalable AL methods have not been investigated yet in RS.

2.4. Aim 4: The availability of increased number of multi-source images (multispectral, hyperspectral and SAR) associated to the same geographical area motivates the need for effective methods, which can fuse the image descriptors to define rich characterization of RS images. However, in RS CBIR scientific literature and existing tools, their fusion has not been explored yet. To better characterize a RS image with multiple modalities, novel and effective methods and tools that are capable of combining information from multiple sensors will be developed. In RS, there are several multi-modal learning algorithms developed for pixel-based single image classification problems, where the aim is land-cover maps generation (i.e., pixel-based image classification problem) [8]. Accordingly, the Aim 4 will be devoted to adopt the state of the art multi-modal learning algorithms (such as multiple kernel learning) to the RS CBIR problems and also develop innovative multi-modal learning methods suitable for the very high spatial/spectral RS image retrieval problems. The Aim 4 will take as input the outcomes (i.e., feature extraction methods) of the Aim 1-2. The novel algorithms will be developed based on Deep Learning (DL) architectures, aiming to learn joint feature representations between different modalities that are invariant across multiple modalities in an unsupervised way. DL has recently attracted great attention in RS due to its effective and accurate feature learning capability [9], whereas its application in such problems has been not investigated yet in RS.

2.5. Aim 5: - During the project, a benchmark archive will be constructed by collecting high amount of freely distributed RS images with multi-modalities (such as: Sentinel-2, Sentinel-1, CHRIS Proba and Hyperion data). High fraction of these images will be annotated by broad category labels and primitive class labels.

4. CONCLUSION

The BigEarth project will address the emerging methods and tools for the accurate and scalable processing of big data in EO. All the developed algorithms will be implemented using open source programming languages and compilers. This will advance the already available tools for RS CBIR problems. All constructed archive (with image annotations) will be made public, and thus the BigEarth project offers to the scientific community the potential to increase the utilization of ever larger archives of EO data. The BigEarth project will allow an efficient discovery of the information existing in the massive EO data archives, i.e., it will provide a very high capability to quickly and accurately access and extract vital information for observing Earth from big EO archives. From an applicative perspective, the developed methodologies and tools are expected to have a very relevant impact on many EO data applications, such as accurate and scalable retrieval of: specific man-made structures, oil slick in the sea, landslides and harmful algal bloom.

3. REFERENCES

- [1] B. Demir, L. Bruzzone "Hashing based scalable remote sensing image search and retrieval in large archives", IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no.2, pp. 892-904, 2016.
- [2] B. Chaudhuri, B. Demir, L. Bruzzone, and S. Chaudhuri, "Region-based retrieval of remote sensing images using anunsupervised graph-theoretic approach," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 7, pp. 987-991, July 2016.
- [3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, 2004.
- [4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. Comput. Vis. Pattern Recognit., 2005, pp. 886-893.
- [5] A. Kaarna and J. Parkkinen, "Comparison of compression methods for multispectral images," in Proc. NORSIG—Nordic Signal Process. Symp., Kolmarden, Sweden, vol. 2, pp. 251-254, 2000.
- [6] Y. Tseng, H. Shih, and P. Hsu, "Hyperspectral image compression using three-dimensional wavelet transformation," in Proc. 21st ACRS, Taipei, Taiwan, 2000.
- [7] G. P. Abousleman, M. W. Marcellin, and B. R. Hunt, "Compression of hyperspectral imagery using the 3-D DCT and hybrid DPCM-DCT," IEEE Trans. Geosci. Remote Sens., vol. 33, no. 1, pp. 26-34, 1995.
- [8] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: a review and future directions", Proceedings of the IEEE, vol. 103, no. 9, pp.1560-1584, 2015.
- [9] A. Romero, and C. Gatta, and G. Camps-Valls, "Unsupervised Deep Feature Extraction for Remote Sensing Image Classification", IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 3, 2016.

NEW METHODOLOGIES TO ANALYZE BIG DATA FROM SPACE WITH A SPECIAL FOCUS ON EARTH OBSERVATION DATA

László Baczárdi, Gergely Bencsik, Zoltán Pödör

Institute of Informatics and Economics, University of Sopron

ABSTRACT

Nowadays, there are lots of data measured by sensors and analyzed by various models and methods. This process is also supported by Big Data environment. Since data processing can be a particularly heavy task, Big Data in space needs special approaches. In our paper, two new methodologies are introduced, which can be adopted to analyze data originated from space with special focus on Earth Observation data. The CReMIT (Cyclic Reverse Moving Intervals Techniques) method extends the analysis possibilities of time series—creating derived, secondary time series—to find more precise correlations. But based on our experiments, these constantly increasing possibilities can cause such conditions near which the results are born just randomly, despite of the exact mathematical environment. Therefore, the random property of the result is hidden from the scientists.

Index Terms— Big Data; CReMIT, Random Correlations, Earth Observation

1. INTRODUCTION

Space research and space activities are one of the most evolving scientific fields. Because of the increasing number of satellites and the numerous observations, the space data volume is getting bigger in the past few years [1]. Space data cannot be called Big Data just because of the data amount (tera-, petabyte scale), but the high velocity (new data is coming continuously and with an increasing rate), the variety of the data (data is delivered by different kind of sensors acting over various frequencies and different aims), as well as the veracity (sensed data is associated with uncertainty and accuracy measurements) support the Big Data properties.

There are several sources including satellites, airborne sensors and in situ measurements which provides wide variety of data. If we can use these data in a combined form, new scientific results can be born. The value of Big Data from space depends on the human capacity too since getting useful information from data originated in space is not trivial. From communication technology point of view, the data must be transferred back to Earth from the satellites [2]. Due of the variety of the data, the data storage can be complicated, because the classic relational databases are optimized to store one pre-defined data types in each column [3]. One of the main characteristics of the Big Data solutions is that they all evolve at a very fast speed [4]. Data are used in different

analysis processes; therefore, the choice of the proper analysis method is also very critical. Different data mining techniques are available for many scientific field to solve their general problems. But specific areas and especially specific problems usually need appropriate algorithm adaptations. The field of Earth Observation is a quickly developed area, and it provides huge number of observed raw Big Data with different properties.

Therefore, new methodologies are needed which are less sensitive for the incoming data types. Our research aim is to introduce the applicability of our newly developed methodologies on data in the domain of Earth Observation.

2. DEVELOPED METHODOLOGIES

It is important to define the adequate methodologies for the treatment of the space datasets. Probably, the above mentioned traditional solutions need special adaptation to reach scientific aims. Two self-developed algorithms, CReMIT and Random Correlations, were applied to get information from Big Data in space environment. By combining CReMIT, Random Correlations and different data mining techniques, more precise results can be achieved.

We would like to remark that these methodologies were developed earlier to examine time series and correlations between different datasets. However, this is the first time that they are applied in a combined way. Both were developed universally and they do not depend of the origin of used data. There are more research areas where we have already used these techniques separately, but they can be extended onto space data.

2.1. The CReMIT method

The collected datasets usually have timestamp, so they are named as time series. Searching for relationships between time series is a major area of statistics and data mining in the domain of space data as well. A huge number of techniques are available, among them correlation, regression analysis and different kinds of data mining techniques are the most frequently used methods for defining connections between one or more independent and dependent variables. Beside the applied methods of analyses, the completeness of the examinations can be significantly affected by the sphere of the involved dependent and independent variables.

If we have a proper length time series, the temporal changes of the relationships using the forward and backward

evolution and moving interval techniques can be examined. The essence of the moving interval technique is that the length of the examined interval is always fixed and the starting point is moved forward by one period in each iteration step. In the case of the evolutionary technique, the starting point is fixed, and the interval length is increased by one period step by step.

Let us given a ts time series: $ts = (s_1; s_2; \dots; s_n)$ with natural period length P . CReMIT uses three user defined parameters: (1) the starting point ($1 \leq SP \leq P$) of the examined time series; (2) the maximum time shifting value (L) and (3) the maximum window width value (J). Based on these parameters and the properties of the examined times series creates all possible derived, secondary time series in the next form:

$$tr_{ts} = \begin{pmatrix} SP + i; SP + i + j \\ SP + i + P; P + i + P + j \\ \vdots \\ SP + i + P * t; SP + i + P * t + j \end{pmatrix},$$

where t is the maximal number of cycle number in the derived, new time series.

$$t = \left\lceil \frac{n-SP-i-j}{P} \right\rceil + 1,$$

where $\lceil \cdot \rceil$ is the entire function. Parameters i and j are the actual time shifting and window width values, $0 \leq i \leq L$ and $0 \leq j \leq J$.

These window-based techniques can be used not only during the breakdown of the whole examined data line into intervals, but also in a more specific manner in the case of periodic time series. A special systematic window concept named as CReMIT was developed which combines the solution of moving intervals and evolutionary techniques [5]. It is important that the CReMIT method is independent of the further applied analysis technique.

We studied how the range of transformation functions can be expanded and how this could be realized. It may mean the application of non-linear functions or some weight function. The application of binary weight factors also allows for the definition of discrete windows in the method.

As a proof-of-concept, the CReMIT method was used on different datasets of forestry and wood science to examine the effect of climate change [6, 7]. In these problems, it was really important to examine the time shifted effect of the climate parameters on the dependent variables, where we did not know the exact measure of time shifting. CReMIT method creates and examines all possible secondary time series. After CReMIT has been applied, we can use all different analyzing methodologies, which are usable on the original time series.

The applicability of CReMIT method is independent of the source of examined data. It requires only the timestamp of the data, therefore the CReMIT method can be use all type of time series to create the secondary datasets for further data analysis. The applicability of the method is not limited to the forestry and wood science problems. It can be applied in any other field, (e.g., on space data) where the analysis of the

relationships between periodic time series has fundamental importance. The application of space data into the relationships examinations of the data from earth proposes many interesting facilities for the future.

2.2. Random Correlations

The main idea behind the theorem of Random Correlations is that data rows as variables present the revealed, methodologically correct results, however, these variables are not always truly connected, and this property is hidden from researchers as well [8].

Data are collected to analyze them and based on the results of analyses, decision alternatives are created. After making the decision, we act according to the selected decision, e.g., we define a correlation between two parameters. All decision processes have a validation phase; however, validation can generally only be done after the related decision has been made. If we have made a false decision, then the consequences lead to a false correlation.

The Random Correlations (RC) theory states that there can be connection between data rows randomly which could be misidentified as a real connection. There are lot of techniques to measure result's endurance, such as R^2 , statistical critical values, etc. The main difference between "endurance measurement" values and RC is the approach of the false result. If we have a good endurance of the result, we strongly assume that the result is fair or the sought correlation exists. RC states that under the given circumstances, we can get results with good endurance. R^2 and critical values can be calculated, decision can be made based on these values, but the result still can be affected by RC. Therefore, it is important to know whether our datasets and the whole analysis have been affected by RC or not.

2.2.1. Parameters of Random Correlations

Every measured data has its own structure. Data items with various, but pre-defined form are inputs for the given analysis. We need to handle all kind of data inputs on the one hand, and to describe all analyzing influencing environment entities on the second hand. For example, if we would like to analyze a data set with regression techniques, then we need the number of points, their x and y coordinates, the number of performed regressions (linear, quadratic, exponential) etc. Having summarized, the random correlation framework parameters are:

- k , which is the number of data columns;
- n , which is the number of data rows;
- r , which is the range of the possible numeric values;
- t , which is the number of methods.

To describe all structure, matrix form has been chosen. Therefore, parameter k , which is the number of data rows [also the columns of the matrix], and n , which is the number of data items contained in the given data row [also the rows of the matrix], are the first two random correlation

parameters. The third parameter, range r means the possible values, which the measured items can take. To store these possibilities, the lower (l) and upper (u) bounds must only be stored. For example, $r(1,5)$ means the lower limit is 1, the upper limit is 5 and the possible values are 1, 2, 3, 4 and 5. Range r is not a very strict condition because the measured values intervals can be defined many times, these values are often between these lower and upper bounds. A trivial way to find these limits, when l is the lowest measured value and u is the highest one. They can be sought non-directly as well. These bounds are determined by an expert in this case. E.g., a tree grows every year, but it is impossible to grow 100 meters from practical point of view. The longitude line is infinity, but it is possible to define l and u . Although in our work integers are used, it is possible to extend this notation for real numbers since the possible continuous nature of the measured data. The continuous form can be approximated with discrete values. In this case, the desired precision related to r can be reached with the defined number of decimals. The sign $r(1,5,')$ means the borders are the same as before, but this range contains all possible values between 1 and 5 up two decimals.

Parameter t is the number of methods. We assume that if we execute more and more methods, the random correlation possibility increases. For example, if $t = 3$, that means 3 different methods are performed after each other to find a correlation.

2.2.2. Models of Random Correlations

Three new models are developed to determine the random factor of the analyses [9]:

1. Ω -model: All possible measurable combinations are produced, and we calculate a rate R , which shows us how many possible data rows cause “correlated” judgement.
2. Θ -method: We determine the chance (rate C) of getting a collision, e.g., finding a correlation. C shows us how many datasets must be measured during the research to get a correlation with at least two datasets for sure.
3. Γ -model: We analyze the subsets of the data pass the given test or not. For all subsets, the given method of analysis is performed. In this case, rate S show us that how many subsets eventuate “correlated” result compare to that subsets which do not.

In the case of the Ω -model, all possible measurable combinations are produced. In other words, all possible n -tuples related to $r(l,u)$ are calculated. Because of parameter r , we have a finite part of the number line, therefore this calculation can be performed. That is why r is necessary in our framework. All possible combinations must be produced, which the researchers could measure during the data collection at all. After producing all tuples, the method of analysis is performed for each tuple. If “correlated” judgment occurs for the given setup, then we increase the count of this “correlated” set S_i by 1. After performing all possible

iterations, the rate R can be calculated by dividing S_i with $|\Omega|$. R can be considered as a measurement of the “random occurring” possibility related to RC parameters. In other words, if R is high, then the possibility of finding a correlation is high with the given method and with the related k, n, r and t . For example, if R is 0.99 , the “non-correlated” judgment can be observed only 1% of the possible combinations. Therefore, finding a correlation has a very high possibility. Contrarily, if R is low, e.g., 0.1 , then the possibility of finding a connection between variables is low.

In the case of the Θ -model, rate C is calculated. This shows how much data are needed to find a correlation with high possibility. Researchers usually have a hypothesis and then they are trying to proof their theory based on data. If one hypothesis is rejected, scientists try another one. In practice, we have a data row A and if this data row does not correlate with another, then more data rows are used to get some kind of connection related to A . The question is how many data rows are needed to find a certain correlation. We seek that number of data rows, after which correlation will be found with high possibility. There is a rule of thumb stating that from 2 in 10 variables (as data rows) correlate at high level of possibility, but we cannot find any proof, it rather is a statement based on experiences. The calculation process can be different depending on the given method of analysis and RC parameters. During the Θ -model calculation process, we generate all possible candidates ($|\Omega|$) based on RC parameters first. We create individual subsets. It is true for each subset that every candidate in the given subset is correlated with each other. We generate candidates after each other and during in one iteration we compare the current generated candidates with all subsets’ all candidates. If we find a correlation between the current candidate and either of the candidates, then the current candidate goes to the proper subset. Otherwise, a new subset is created with one element, i.e., with the current candidate. C is the number of subsets. C shows us that how many datasets must be measured during the research to get a correlation with at least two datasets for sure. Based on value C , we have three possible judgements:

- C is high. Based on the given RC parameters, it must be lots of dataset to get a correlation with high probability. This is the best result, because the chance of RC is low.
- C is fair. The RC impact factor is medium.
- C is low. The worst case. Relatively few datasets can produce good correlation.

Using Γ -model, all subsets of the given data items are produced. For all subsets, the given method of analysis is performed. In this case, rate S show us that how many subsets eventuate “correlated” result compare to that subsets which do not.

3. APPLYING CREMIT AND RANDOM CORRELATIONS FOR BIG DATA IN SPACE

To achieve scientific results from row measurement data, three main phases can be defined: (1) data collection, (2) data analysis and (3) validation.

CreMIT method is related to (2), while Random Correlations to (3). CReMIT can systematically produce individual data rows derived from the original data set. Therefore, several “new” data rows are created, and this property extends the analysis possibilities of the classic methods, seasonality and time-shifted correlations can be revealed. Following the application of CReMIT, using Random Correlations methodology ensures that the given results are not just born randomly, i.e., because of the increased number of data rows.

Since CReMIT and Random Correlations do not depend on the input data, they can be used for space and Earth Observation data as well. First, CReMIT are used on the measured data set to create several “new” derived data rows. Then the desired method (method of analysis), such as regression methods or data mining techniques, is performed not just the original data set, but on each “new” data row. Then, one of the Random Correlations models can be applied to determine the random impact level of the given result. If we assume, that the result space is not balanced, then rate R can be calculated. If we would like to determine how many data row should be produced to get a good regression, Θ -method can be applied. With calculating rate S , the stability of the given data row can be determined.

4. CONCLUSION

Our two new, self-developed methodologies can be applied on big datasets in the space domain. Both CReMIT and Random Correlations extend the analysis possibilities of the original data which lead to new results.

The CReMIT method combines the evolution technique and the moving interval technique. Scientists can define the maximum time shift and the maximum window size, then the CReMIT creates all possible derived time series inherited from the original time series. Accordingly, these new time series, the examination possibilities are spread out without reference to the further used analysis methods. The Random Correlations method can determine the random factor of a given analysis. It is important to remark that we do not deny that real connections exist, but we introduced such analysis environments which can examine spurious correlations. Since this kind of correlation can misidentify as real correlations, it is recommended to always calculate how big the possibility of Random Correlations can be and to analyze whether the results space balanced or not.

So far, these new methodologies have been used on datasets from the domain of forestry and wood science, so we extended them to be used in space domain. In our extension, we focused on Earth Observation datasets.

Acknowledgement

The research has been supported by the UNKP-17-4-III New National Excellence Program of the Ministry of Human Capacities.

5. REFERENCES

- [1] C. Arviset, et al., “Big Data, Big Data Challenges And New Paradigm For The Gaia Archive,” *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*, pp. 9-12, 2016.
- [2] M. Pritchard and J. Churchill, “Enabling High Performance Access To Big Data From Space,” *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*, pp. 21-22, 2016.
- [3] S. Natali, M. Folegani and S. Mantovani, “The (Slow) Migration From Image-Based Information Extraction To Data Stream Information Extraction,” *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*, pp. 46-49, 2016.
- [4] R. Vitulli, P. Armbruster and D. Merodio-Codinachs, “Big Data Starts On-Board,” *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*, pp. 141-144, 2016.
- [5] Z. Pödör, M. Edelényi and L. Jereb, “Systematic Analysis Of Time Series – CReMIT,” *Infocommunications Journal 6 (1)*, pp. 16-21, 2014.
- [6] E. Führer, M. Edelényi, L. Horváth, A. Jagodics, L. Jereb, Z. Kern, A. Moring, I. Szabados and Z. Pödör, “Effect Of Weather Conditions On Annual And Intra-annual Basal Area Increments Of A Beech Stand In The Sopron Mountains In Hungary,” *Idojaras, Journal of the Hungarian Meteorological Service 120 (2)*, pp. 127-161, 2016.
- [7] Gy. Csóka, Z. Pödör, Gy. Nagy and A. Hirka, “Canopy Recovery Of Pedunculate Oak, Turkey Oak And Beech Trees After Severe Defoliation By Gypsy Moth (*Lymantria Dispar*): Case study from Western Hungary,” *Lssnicky Casopis - Forestry Journal 61*, pp. 143-148, 2015.
- [8] G. Bencsik and L. Bacsárdi, “Novel Methods For Analyzing Random Effects On ANOVA And Regression Techniques,” *Advances in Intelligent Systems and Computing 416*, Springer, pp. 499-509, 2016.
- [9] G. Bencsik, “Decision Support And Its Relationship With The Random Correlation Phenomenon,” *Ph.D. Dissertation*, 2016.

BIG DATA VISUALIZATION TOOLS IN EO MOBILE APPS

C. Orrù¹, J. Balhar², A. Stoica³, P. Sacramento⁴, G. Rivolta¹

¹Progressive Systems Srl, ²Gisat s.r.o., ³Terrasigna, ⁴Solenix GmbH

ABSTRACT

The Copernicus Sentinel App and My Vegetation App are two mobile applications developed to satisfy the needs of different user profiles, with a variety of objectives ranging from educational and outreach purposes as well as for several communities which include researchers but also general public. This paper presents the criteria used to perform data reduction aimed at (Big) data visualization and the main features of these mobile apps showing which solutions and tools have been adopted for the extraction and visualization of big amount of data.

Index Terms— Visualization, Big Data, RasDaMan, Web World Wind

1. INTRODUCTION

The Copernicus Sentinel App (formerly known as Sentinel App) and the My Vegetation App are two mobile applications developed for ESA within the Fast Prototyping Frame Contract*. The main objective for the App development was to increase the awareness of the general public as well as of potential new users, about the Earth Observation (EO). Such objective is achieved through the implementation of attractive mobile apps and prototypes to experiment different technical solutions and concepts.

The Copernicus Sentinel App is a mobile application developed for both Android [1] and iOS [2] devices and it is available for download via Apple and Google Stores.

The My Vegetation App is the evolution of the Proba-V App. Also this App has been developed for both iOS [3] and Android [4] devices and it is available for download via Apple and Google Store. In the past months the scope of the App has been extended in order to keep promoting the Proba-V mission and products but also to include new features and additional vegetation indexes and parameters delivered by the Copernicus Global Land Service.

These Apps will be further evolved in the "Maintenance and evolution of the Copernicus Apps for mobile devices" contract† that has been recently kicked-off. All new releases will have the Copernicus branding.

* ESA Contract No. 4000112250/14/I-NB, Consortium: Solenix Deutschland GmbH (Prime Contractor), Qualteh JR SRL, Terrasigna SRL, GISAT SRO, Progressive Systems SRL

† ESA Contract No. 4000121771/17/I-SBo, Consortium: Solenix Deutschland GmbH (Prime Contractor), Qualteh JR SRL, Terrasigna SRL, GISAT SRO, Progressive Systems SRL

In the next sections, the details of the criteria used to perform the data reduction necessary to visualize (Big) data in the context of the apps, the choices performed and the tools used for the visualization of big amount of data will be described.

2. COPERNICUS SENTINEL APP

2.1. Main features

The Copernicus Sentinel App aims at providing a simple interface allowing access to different information about the Copernicus Sentinel missions and their satellites.

The main features of the Copernicus Sentinel App are:

- Track in real time the Copernicus Sentinels' satellite position around the 3D globe model of the Earth
- Retrieve information about the last data downlink of the satellite towards the Ground Stations and the last and the next passes of the Copernicus Sentinels over the user position
- Visualize the footprints of the Copernicus Sentinel satellites' data acquisitions in a selected range of time of interest, getting detailed info for each of them
- Have an overview of the possible applications of the Copernicus Sentinels thanks to a selected set of use cases
- Visualize and interact with the 3D models of the Copernicus Sentinels satellites
- Visualize density maps of the big amounts of products of the Copernicus Sentinels satellites
- Get more precise info by increasing the level of zoom
- Compare the density of products of different sensors or satellites (see Fig. 1)
- Select different periods of interest and visualize the related density maps

2.2. Big Data visualization in the Copernicus Sentinel App

After having described the main features of the Copernicus Sentinel App it is easy to understand that the part

related to the density maps of products is the most interesting in terms of Big Data representation.

The adopted solutions aim to aggregate information and visualize it in a clear and efficient way for the user, even if he is not an expert in the Earth Observation field.

Large volumes of data acquired by different missions and sensors are visualized both through 2D and 3D maps in different levels of detail depending on the zoom level.

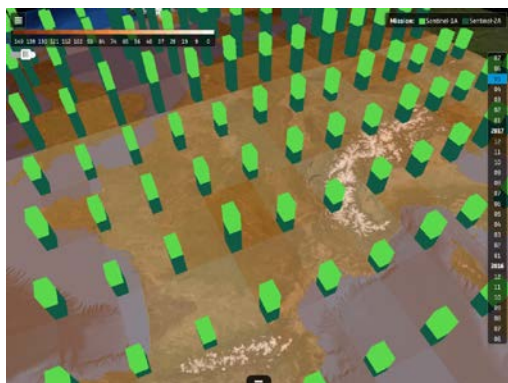


Fig. 1 Density of Sentinel-1A and Sentinel-2A products over France

To create 2D density maps, the choropleth grid [5] has been chosen. This grid allows to easily display the following aggregated data in different levels of detail:

- Total for multiple missions or single mission
- Total for multiple types of instrument or single instrument

This method is based on the division of the area in same (or similar) geometric objects. In particular, the earth is divided into $1^\circ \times 1^\circ$ rectangles in the most detailed resolution. To ensure a good user experience, the data are first pre-aggregated to this minimum grid. Reducing the zoom level, the data are aggregated on the server side by multiple adapters, which are the core part of the data reduction.

In the colour palette, different intensities of colours are associated to different amounts of data.

To display the contributions made by different missions and different instruments, the extruded polygons are used on top of the choropleth grid. The extruded polygons are in fact 3D summarizing diagrams (see Fig. 1).

Each diagram represents the total amount of the missions' products or instrument products acquired in the given area. Different colours are used for distinguishing the particular missions/instruments.

To display the information over the globe, the WebWorldWind (WWW) framework [6], an open-source virtual globe solution, is used. WWW is a library and API, which is able to support Big Data visualization and it is also very well optimized to work on mobile as well as in desktop browsers.

The information about the products is provided by the ESA RSS service [7], which supplies aggregated data for Copernicus as well as for other ESA and Third-Party missions.

Moreover, additional OpenSearch data resources - such as the Federated EO catalogue (FedEO) [8] and the Copernicus Sentinel Scientific Data Hub (SciHub) [9] - could be used to retrieve the information about the products and aggregate them in geographical and temporal grids.

3. MY VEGETATION APP

3.1. Main features

The My Vegetation App is the follow-up of the Proba-V App, which visualizes data from the PROBA-V Mission Exploitation Platform (MEP) [10].

The main additional features of the My Vegetation App are:

- Take a picture and retrieve the values of the following vegetation parameters: Normalized Difference Vegetation Index (NDVI), Leaf Area Index (LAI), Dry Matter Productivity (DMP) and Land Surface Temperature (LST) related to the position of the user
- Access to the graphic of the time series of the vegetation parameters values (see FIG. 2)
- Share and edit the taken picture on social media
- Visualize over a 2D map the following products: Corine Land Cover, NDVI, LAI and Natura 2000 areas
- Select any point in the map and retrieve the graphics of the time series of the vegetation parameters values
- Compare the time series of the vegetation parameter values related to two different map locations

3.2. Big Data visualization in the My Vegetation App

As can be inferred from the previous section, the My Vegetation App supports the visualization of time series extracted from big amounts of data.

Beyond the service provided by the PROBA-V MEP, RasDaMan (Raster Data Manager) [11] [12] is used as a temporary Database Management System for storing and delivering vegetation indexes. The App performs HTTP requests (Latitude, Longitude and Time interval) to the server hosting the RasDaMan instance, which returns a response in XML format.

These server responses are then processed and the values are displayed in the App when the user takes a picture or when he taps on the map. Global rasters (matrixes covering the globe) are imported automatically from land.copernicus.vgt.vito.be into RasDaMan time-series cubes for NDVI, LST, DMP and LAI.

The RasDaMan data stored for more than 1 year is reaching the size of approximately 1 TB.



FIG. 2 Time Series of LAI Index

Data used in the MyVegetation app represent EO Level 3 products systematically derived from Proba V mission acquisitions. Such vegetation related data are available via the app for visualization over the entire globe, ranging from coarse to medium spatial resolution, and currently cover a period of 3 years.

Although the volume of stored data in the RasDaMan is at the moment around 1 TB (increasing every day), the corresponding number of L3 values available for visualization is huge, and the possibility to search and retrieve information from a huge number of stored values and display them in time series or single values is relevant in the Big Data context.

For the MyVegetation app, such feature enables access to vegetation information also in isolated places, in nature, where there is no computer infrastructure available, allowing interested users to find out real and up to date vegetation information related to any remote location and share it within their social networks.

In general, the possibility of displaying different EO data products is a valuable feature, potentially enabling mobile apps to be effective means for EO outreach and promotion.

4. CONCLUSIONS

The main features and functionalities of the Copernicus Sentinel App and the MyVegetation App have been presented in this paper.

Tools that can be used in mobile applications for Big Data visualization, helping the general public to easily understand the Earth Observation (EO) world and increasing the user uptake of EO data, have been demonstrated.

5. REFERENCES

- [1] <https://play.google.com/store/apps/details?id=esa.sentinel&hl=en>
- [2] <https://itunes.apple.com/us/app/esa-sentinel/id1036738151?mt=8>
- [3] <https://itunes.apple.com/us/app/esa-proba/id1098681425?mt=8>
- [4] <https://play.google.com/store/apps/details?id=eoapps.probava&hl=en>
- [5] T. Giraud, N. Lambert, "Reproducible Cartography", International Cartographic Conference ICACI 2017: Advances in Cartography and GIScience pp 173-183, 2017
- [6] Y. Voumard, P. Sacramento, P. G. Marchetti, P. Hogan, WebWorldWind, achievements and future of the ESA-NASA partnership, PeerJ Preprints | <https://doi.org/10.7287/peerj.preprints.2134v2> | CC BY 4.0 Open Access | rec: 30 Sep 2016, publ: 30 Sep 2016
- [7] P.G. Marchetti, G. Rivolta, S. D'Elia, J. Farres, G. Mason and N. Gobron, "A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support", IEEE Geoscience and Remote Sensing (162): 10-18, 2012
- [8] <http://fedeo.esa.int/opensearch/readme.html>
- [9] <https://scihub.copernicus.eu>
- [10] E. Goor et al. PROBA-V "Mission Exploitation Platform Remote Sens". 2016, 8(7), 564; doi:10.3390/rs8070564
- [11] P. Baumann, "rasdaman: Array Databases Boost Spatio-Temporal Analytics", Computing for Geospatial Research and Application (COM.Geo), 2014 Fifth International Conference on, IEEE, 2014, DOI: 10.1109/COM.Geo.2014.1
- [12] P. Baumann, "The DataCube Manifesto" http://earthserver.eu/sites/default/files/upload_by_users/The-Datacube-Manifesto.pdf, accessed on 3.08.2017

BIG LUNAR DATA VISUALIZATION AND ANALYSIS

Emily Law¹, George Chang¹, Richard Kim¹, Shan Malhotra¹

(1) Jet Propulsion Laboratory, California Institute of Technology (Emily.S.Law@jpl.nasa.gov, George.Chang@jpl.nasa.gov, Richard.M.Kim@jpl.nasa.gov, Shan.Malhotra@jpl.nasa.gov)

ABSTRACT

NASA's earth and planetary spacecraft return large amounts of remote sensing data, such as imagery and raw science measurements, in support of remarkable research. Not only does the data lead to new scientific discoveries about our planet and the solar system, it provides a wealth of information to educate, inspire, and engage the public at large. To leverage this rich data for mission planning, scientific research, public outreach and education, it is essential to make it accessible and understandable, analyzable, all while appealing to their interests. This presentation will highlight web-based capabilities that showcase NASA's large volume of lunar data collected from past and current Moon missions. It is particularly relevant as the new Administration has more plans for the Moon. We will illustrate big data visualization and analysis in easy-to-use and interactive mediums for diverse use.

1. INTRODUCTION

Meeting the challenges of space exploration has resulted in new knowledge that has kept NASA and JPL world leaders in big data science and technology. We, NASA's Solar System Treks project development team at JPL, under the management of Solar System Exploration Virtual Research Institute (SSERVI) [1], have developed a system that includes a set of web-based portals and a suite of interactive visualization and analysis tools to enable mission planners, scientists, and engineers to access a large volume of mapped data products and models from past and current missions. Currently, three web-based portals are publicly available for discovering the Moon, Mars and Vesta. This presentation will provide an overview of this system highlighting its lunar web portal, Moon Trek (<https://moontrek.jpl.nasa.gov>) that was designed and developed specifically for exploration of our Earth's Moon. We will also demonstrate its uses, features and capabilities, highlighting big lunar data visualization and analysis innovations.

2. MOON TREK

Moon Trek provides a suite of interactive tools that incorporate observations from past and current lunar

missions, creating a comprehensive lunar research and educational web portal. The online web portal allows anyone with access to a computer to search through and view a vast number of lunar images and other digital products without having to download any application to the computer. The portal provides easy-to-use tools for browsing and searching, data layering and feature search, including detailed information on the source of each assembled data product from NASA's Planetary Data System (PDS) [2]. While mission planning was initially the primary emphasis when our work started in the NASA Constellation Program era, Moon Trek has evolved and expanded to address the lunar science community, the lunar commercial community, education and outreach, and anyone else interested in accessing or utilizing lunar data. Its visualization and analysis tools allow users to perform analysis of big volumes of lunar data such as lighting and local hazard assessments including slope, surface roughness and crater/boulder distribution. Moon Trek features a generalized suite of tools facilitating a wide range of activities including the planning, design, development, test and operations associated with lunar sortie missions. Sharing of multi-layered visualizations is made easy with the ability to create and send using URL-encoded links. Moon Trek is also a powerful tool for education and outreach, as is exemplified by being designated as a key supporting infrastructure for NASA Science Mission Directorate's STEM Activation Initiative, as data service to NASA's Eyes on the Solar System, and as serving of data to a growing community of digital planetariums.

3. BIG DATA

Large amount of lunar data from the Apollo era to the latest instruments (such as the high definition camera) on board the Lunar Reconnaissance Orbiter (LRO) have been collected by NASA and other international space agencies. Although raw data/images and calibrated data sets are accessible via PDS, additional processing is required to transform these raw data and images to geo-referenced, aggregated and mosaicked images in order for Geographical Information System (GIS) like Moon Trek to display them as rich visualization layers that are highly valuable for future mission planning and development, scientific research as well as for general public. Thousands of such

processed higher level visualization data products that are map projected, georeferenced, ortho-rectified and controlled under the same Lunar Orbiter Laser Altimeter (LOLA) network are made available through the Moon Trek. These large amounts of data Moon Trek manages ranges from a few gigabytes to hundreds of gigabytes in size with new data products adding to the system when they are available. One example of the data Moon Trek serves is shown in Figure 1. It is a close-up image taken by the narrow angle camera aboard the Lunar Reconnaissance Orbiter. The image is (only a sub-image is shown) has a resolution of 5064x52224 (264 megapixels) as a 264 megabyte TIFF image. The physical resolution of the image is approximately 1 meter/pixel. We currently have thousands of these images.



Figure 1. LRO Narrow Angle Camera Image

Another example shown in Figure 2 is a composite image consisting of multiple images taken by Apollo 15 Metric Camera imaging system as it orbited the Moon. The size of the image is 49152x33496 (1.6 gigapixels) as a 3.3 gigabyte TIFF image. The physical resolution of the image is approximately 10 meters/pixel.

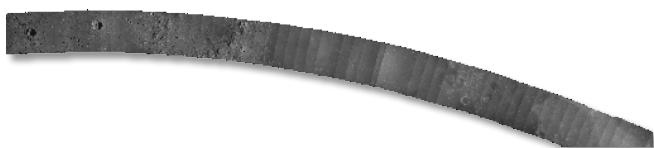


Figure 2. Apollo 15 Orbit 33 Mosaic

Figure 3 shows a complete lunar map composed of many regional images collected by the Ultraviolet/Visible camera (UVVis) aboard the Clementine spacecraft. The image size is 92160x46080 (4 gigapixels) compressed with JPEG2000 to 200 megabytes with a physical resolution of approximately 100 meters/pixel.

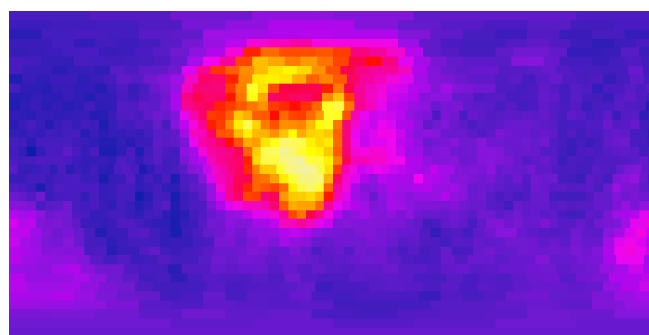


Figure 3. Potassium Concentration Map

Despite this ever-increasing amount of data, Moon Trek must provide users with the best performance for browsing, viewing and analyzing the vast amounts of data available in a timely manner.

4. SYSTEM TECHNOLOGIES

Moon Trek as well as other web portals provided by the Solar System Treks project (e.g., Mars Trek <https://marstrek.jpl.nasa.gov>) is built on a Service Oriented Architecture (SOA) [3] that is scalable and extensible for all planetary bodies. All Trek portals are supported by a common backend infrastructure and use a common frontend visualization framework. Figure 4 below depicts the high level architecture of the Solar System Treks system that serves as the foundation of the Moon Trek (as well as other Trek portals). The infrastructure provides core services for data ingestion, data management, image tiling using hadoop [4], arcGIS [5], and workflow using cloud computing for various computation and data analysis services, search via SOLR [6] and download. By taking advantage of Amazon Cloud Front [7], the system securely and effectively delivers Moon Trek data products to the front end. It provides OGC [8] compliant standard web services APIs for accessing all data products. The frontend framework takes advantage of HTML5 [9] frontend user interface. It employs a flat open space design and implementation that maximizes usability with modular tools and widgets. It includes in browser over sampling that stretches the image dynamically when zoom in past its resolution level. It is responsive to different sizes and form factors. It is embeddable into other browsers. It employs Cesium [10], a JavaScript library that supports 3D globe view for visualizing dynamic data with high rendering performance, precision, visual quality and ease of use. By using standard keyboard gaming controls, Moon Trek allows users to maneuver a first-person visualization of “flying” across the surface of the Moon. Users can also specify any area of interest to generate STL [11] or OBJ [12] files for creation of physical models of surface features with 3D printers.

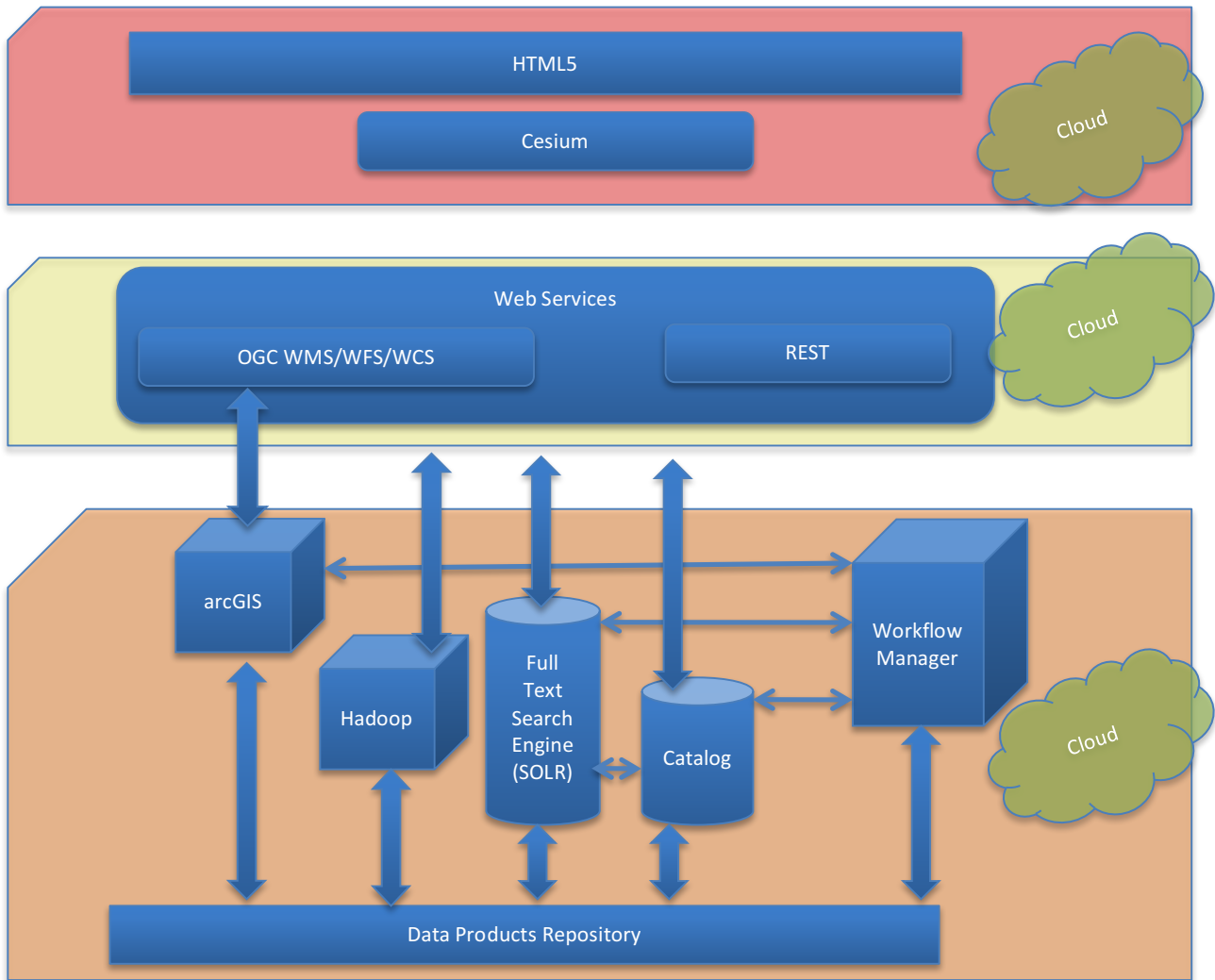


Figure 4. High Level Architecture

5. SUMMARY AND CONCLUSION

NASA’s Solar System Treks project has grown considerably from its initial goal of providing a mission planning tool for Lunar exploration. Recognizing the big data trend and challenges, the team had the foresight to architect and design the system (including a combination of a backend infrastructure and a common user interface framework) to be scalable and extensible. The system is now successfully providing capabilities for multiple activities including mission planning, scientific research, decision making, as well as public outreach for other planetary bodies via its web portals. In addition to continual enhancements to the system, the team continues to expand its capabilities to new destinations and new research. We also encourage and invite

the user community to provide suggestions and feedbacks as the development team continues to expand the capabilities of the system, its related products, and the range of data and tools that we have provided.

6. ACKNOWLEDGEMENT

The authors and the Solar System Treks team would like to thank the Advanced Explorations Systems Program of NASA's Human Exploration Operations Directorate, the Planetary Science Division of NASA's Science Mission Directorate, and the Solar System Exploration Virtual Research Institute for their support and guidance in the continuing development of Solar System Trek portals.

7. REFERENCES

- [1] SSERVI <https://sservi.nasa.gov/>
- [2] PDS <https://pds.nasa.gov>
- [3] "What Is SOA?".
<https://www.opengroup.org/soa/source-book/soa/soa.htm>
- [4] Hadoop "Welcome to Apache Hadoop!".
hadoop.apache.org
- [5] ArcGIS <https://www.arcgis.com/features/>
- [6] SOLR <http://lucene.apache.org/solr/>
- [7] Amazon Cloud Front
<https://aws.amazon.com/cloudfront/>
- [8] OGC <http://www.opengeospatial.org>
- [9] HTML5 Leslie Sikos. "HTML5 Became a Standard, HTML 5.1 and HTML 5.2 on the Way"
- [10] Cesium <https://cesiumjs.org>
- [11] STL [https://en.wikipedia.org/wiki/STL_\(file_format\)](https://en.wikipedia.org/wiki/STL_(file_format))
- [12] OBJ https://en.wikipedia.org/wiki/Wavefront_.obj_file

THE REVISED TIME-FREQUENCY ANALYSIS (R-TFA) TOOL OF THE SWARM MISSION

Balasis, Georgios⁽¹⁾, Papadimitriou, Constantinos⁽¹⁾, Daglis, Athanassios⁽¹⁾, Giannakis, Omiros⁽¹⁾, Giamini, Sigiava A.⁽¹⁾, and Vasalos, Georgios⁽¹⁾

¹National Observatory of Athens, IAASARS, Greece

ABSTRACT

The IAASARS/NOA team has developed versatile signal processing tools suitable for the verification and validation of Swarm Level 1b products of the electric and magnetic field. These tools can be used for the study of magnetospheric Ultra Low Frequency (ULF) waves and ionospheric Equatorial Spread-F (ESF) events, which are phenomena of space weather with effects on technology infrastructure. They are also useful for deriving sets of data suitable for lithospheric and main magnetic field modelling. These tools have the capability to extract information needed to detect space weather events as well as to complement other validation tools used on Swarm mission. Developed to derive the characteristics of ULF waves, the Time-Frequency Analysis (TFA) methodology, based on wavelet transforms, has proven to be effective when applied to the Swarm data to retrieve, on an operational basis, new information about the near Earth electromagnetic environment. Here, we present the Revised TFA (R-TFA) tool introducing a new user friendly "web" interface. The main advantage of the R-TFA tool is that is available publicly via a web browser to the scientific community without any software/hardware requirements and tedious installation processes.

Index Terms— Earth Observation, Swarm mission, Time series visualization, Time series analysis, Space Weather

1. INTRODUCTION

The IAASARS/NOA team has designed and developed an interactive tool for the exploitation, selection and visualization of data from ESA's Swarm mission using state-of-the-art visualization libraries and implemented proper search tools integrated with a single MySQL database.

Swarm is the fourth Earth Explorer mission in ESA's Living Planet Program launched on November 22, 2013. The database includes specific constellation spacecraft instruments data of interest (primary dataset) and other data/metadata (secondary datasets) of scientific association to the primary ones. The interface between the databases and the outer layer of the tool are implemented through a set of php routines prepared in accordance to the user-needs specification (Figure 1). These routines facilitate the data access for the subsequent data selection, visualization, exploitation and retrieval. The users may also have the capability of building and applying dedicated combined search tools in order to exploit the primary datasets using available functions and routines. Moreover, combined searches based on the values, the duration and/or basic statistical properties of the primary and/or secondary data series are provided. Standard search tools based on mission selected times and/or selected mission segments are used as basic or additional search filters.

The web-based tool makes use of the time-frequency analysis (TFA) tool developed within the frame of ESA study "Multisatellite, multi-instrument and ground-based

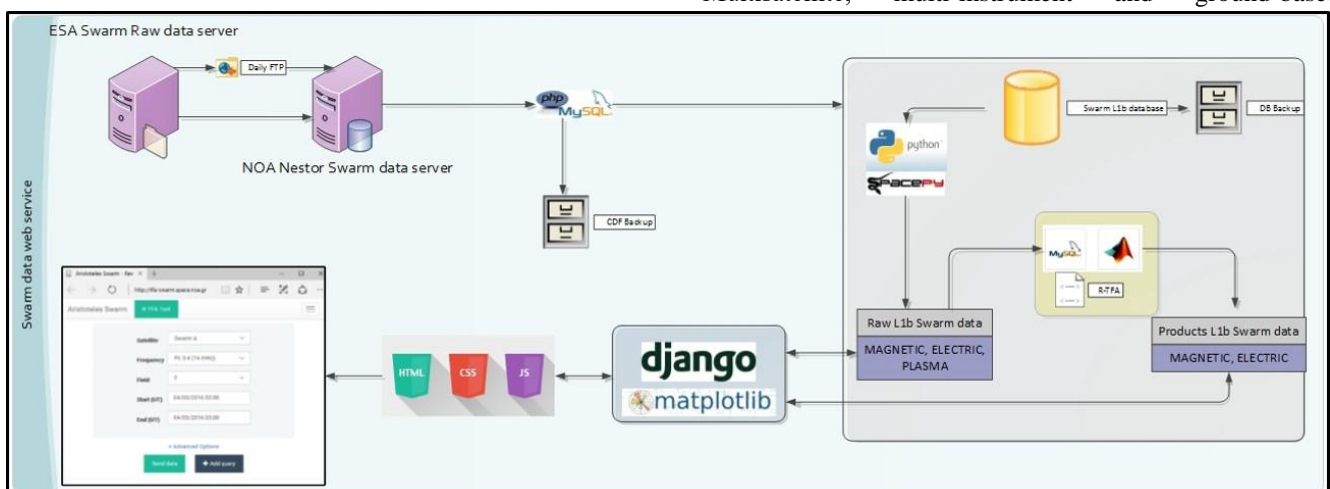


Figure 1: R-TFA tool flowchart

observations analysis and study of ULF wave phenomena and products (ULFwave)” (Balasis et al., 2013). The TFA tool is an advanced suite of algorithms based on wavelet transforms, tailored to the analysis of Level 1b data from the Swarm mission (Balasis et al., 2015). The aim of the TFA tool has been to combine the advantages of multi-spacecraft and multi-instrument monitoring of the geospace environment in order to analyze and study magnetospheric ULF waves (Balasis et al., 2016).

2. R-TFA TOOL WEB-BASED SERVICE

Visualization interactive tools are available for the datasets ingested in the database to facilitate further data exploitation. The selection of data can be based on temporal or spatial criteria, according to the satellites’ position and additionally, since Swarm is a multi-satellite mission, according to the satellites’ relative position/orientation and flight configuration.

Moreover, metadata of Swarm interest, such as the values of various geomagnetic indices, like the Kp, Dst or Ap, are also ingested in this tool in order to facilitate the extraction of Swarm data during periods of enhanced geomagnetic activity. The user has the option to select Swarm measurements in comparison to similar measurements obtained by another ESA mission, for instance Cluster magnetospheric mission, when the two missions are in local time conjunction. On top of our development was to retain the fast, dynamic and interactive character required to modern visualization tools, which are motivated from the complexity, scale and variety of satellite data, the plethora of data providers, the wide variety of formats and data policies, as well as the uncertainty of data quality.

IAASARS/NOA visualization tool is fully interactive and provides an exploratory to help generate ideas, easily accessible through an easy to use (web-based) data management platform, integrative for bringing multiple data products or parameters together, informative and engaging.

Figure 2: Web-based user-friendly R-TFA query forms

Furthermore, data from a specific instrument can be selected according to criteria based on measurements performed by other instruments, i.e. intervals of magnetic field data can be selected based on when the satellites’ electric field instrument records exhibit values within a specific range, as well as criteria based on the measurements performed by one of the other satellites of the constellation, or even of the properties of their collective measurements, i.e., when mean values across all three satellites exceed a particular threshold, giving an indication for the general conditions of their environment.

The dynamic character of Swarm visualization tool, allow users to gain insight to multi-scale data preview (changing from global to local), preview large data holdings, inter-compare diverse product parameters, be able to assess different product quality and make data accessible to new audiences. Dynamic and interactive time series and horizon graphs can be easily created from the reference data/metadata, providing simple or complex – depending on the user’s selection criteria and queries – data-driven visualization capabilities. As stated previously, time intervals

can be selected and changed at will in a fast and robust manner, zooming capabilities are ubiquitous and different tools provide additional options such as dynamic tooltips, through which the user can have access to additional point by point information that would have been extremely

cumbersome or even impossible to get through the options provided by more traditional approaches.

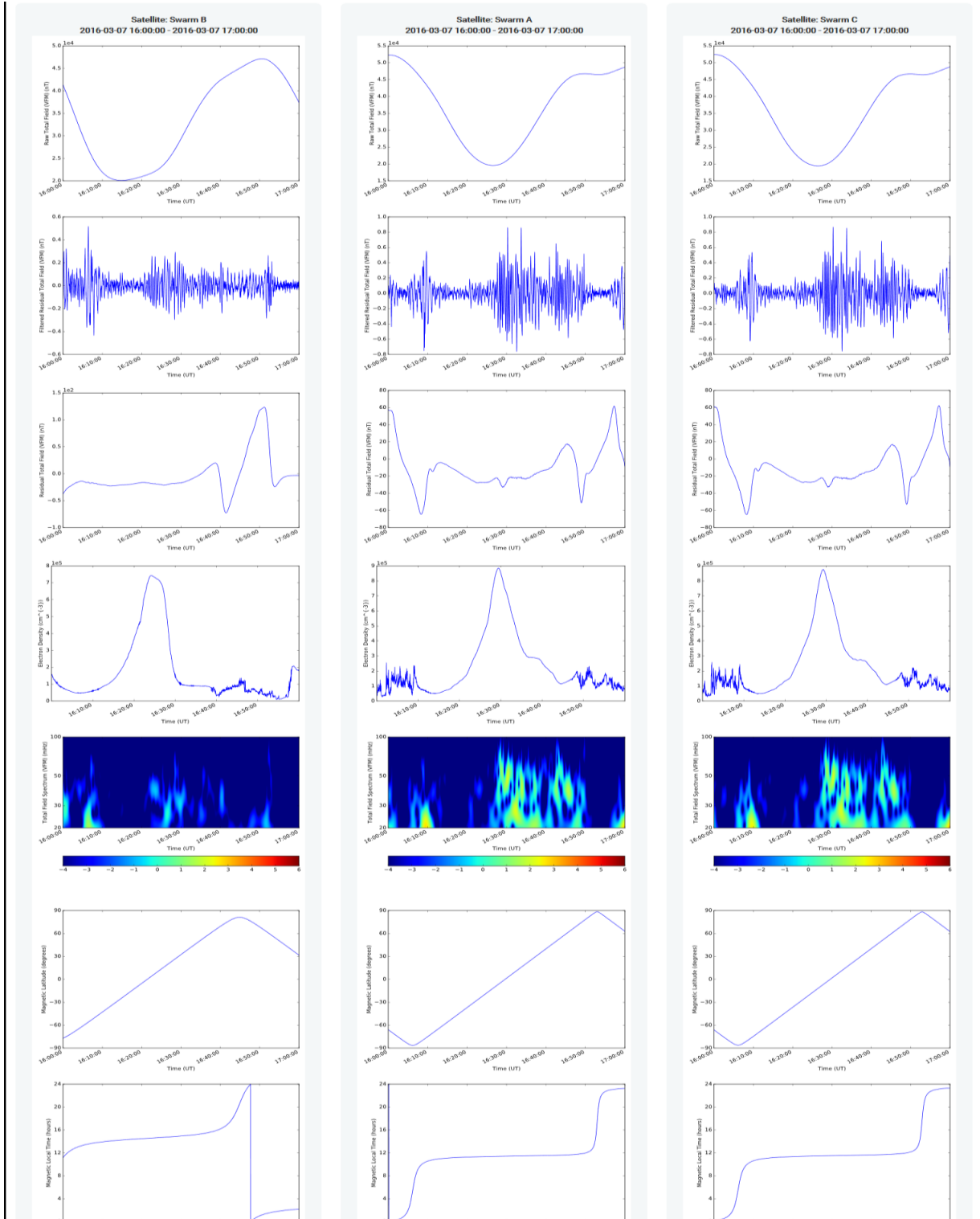


Figure 3: Output visualizations of the R-TFA tool

3. SUMMARY

The goal of the revised TFA (R-TFA) tool is to equip the scientific community with an easy to use, web-based application that will provide simple and intuitive visualizations of both raw data and derived products as listed in Figure 2. By using HTML5, CSS3, Bootstrap and JavaScript, we have created a user-friendly web User Interface that will be accessible from everywhere. The R-TFA-tool consists of two main graphic user interfaces (GUIs). In the first part (Figure 2), the user has the option to choose from a number of datasets. Data from low-Earth orbit (LEO) satellite missions like Swarm, CHAMP and ST5 are available as well as data from ground-based magnetometer networks such as the Hellenic GeoMagnetic Array (ENIGMA) [<http://enigma.space.noa.gr/>]. For each dataset, the GUI automatically loads the corresponding options in drop-down menus like the satellite, the type of measurements (high/low resolution) and the type of measured field magnetic/electric or particles) as well as one of their relevant components (X/Y/Z/Total). Users can also choose the time range that they wish to examine along with the range of frequencies of ULF waves. The tool also provides the capability to plot additional time series that might be useful for interpreting the results of the analysis, like the time series at the different pre-processing stages (raw data / residual / filtered), along with other, complementary parameters, such as the satellite's position (in both geodetic and magnetic coordinate systems), data from the Swarm level 2 products (Field Aligned Currents – FAC / Ionospheric Bubble Index - IBI) and the values of various indices of geomagnetic activity (Kp, Dst). After all the user choices are made the data is retrieved and the second (plotting) GUI is activated (Figure 3). In this GUI, the user can interactively view measurements and pertaining data in easy-to-inspect organized plots and browse through the results in a track-by-track fashion.

4. ACKNOWLEDGEMENTS

This work was supported by the Excellence Research Program GSRT (2015-2017) ARISTOTELIS "Environment, Space and Geodynamics/Seismology 2015-2017".

5. REFERENCES

- [1] Balasis, G., C. Papadimitriou, E. Zesta, and V. Pilipenko, *Monitoring ULF Waves from Low Earth Orbit Satellites*, in *Waves, Particles, and Storms in Geospace*, ed. by G. Balasis, I. A. Daglis, and I. R. Mann, 347 – 352, Oxford University Press, 148–169, 2016
- [2] Balasis, G., C. Papadimitriou, I. A. Daglis, and V. Pilipenko, "ULF wave power features in the topside ionosphere revealed by Swarm observations", *Geophys. Res. Lett.*, 42, 6922–6930, doi:10.1002/2015GL065424, 2015
- [3] Balasis, G., I. A. Daglis, M. Georgiou, C. Papadimitriou, and R. Haagmans, "Magnetospheric ULF wave studies in the frame of Swarm mission: A time-frequency analysis tool for automated

detection of pulsations in magnetic and electric field observations", *Earth Planets Space*, 65, 1385–1398, 2013

- [4] Stolle, C., H. Lühr, M. Rother, and G. Balasis, "Magnetic signatures of equatorial spread F as observed by the CHAMP satellite", *J. Geophys. Res.*, 111, A02304, doi:10.1029/2005JA011184, 2006

BIG DATA ANALYTICS APPROACH FOR GEOSPATIAL INVESTIGATION OF URBAN GEOHAZARDS IN NAPLES, ITALY, WITH COSMO-SKYMED PERSISTENT SCATTERERS

Tapete D.¹, Cigna F.¹, Milillo P.², Perissin D.³, Serio C.⁴, Milillo G.¹

¹ Italian Space Agency (ASI), Italy; ² NASA-JPL, USA; ³ Purdue University, USA;

⁴ University of Basilicata, Italy

ABSTRACT

A tiered data analytics approach is presented to exploit big COSMO-SkyMed Persistent Scatterers time series, at the finest level to identify urban geohazards. The Metropolitan City of Naples is discussed as an example of dynamic and complex urban environment where motions estimated from satellite Interferometric Synthetic Aperture Radar (InSAR) can be used as proxies of warning, if captured and creamed off from redundant data.

Index Terms— Big Data, InSAR, COSMO-SkyMed, data analytics, urban remote sensing

1. INTRODUCTION

In the last years satellite Interferometric Synthetic Aperture Radar (InSAR) has been increasingly used to sense and monitor dynamics of large and developing cities in Italy. In particular, Persistent Scatterer (PS) InSAR proved valuable for urban geohazard assessment and early warning, when used in integration with other informative layers to correlate observed motions with local geological setting, exploitation of geo-resources, infrastructure development, and condition and use of pre-existing urban structures [1-2].

However, there are two factors that urban remote sensing scientists need to account for:

- (i) the complexity of urban environments and the dynamicity of the processes transforming them, which make the correct attribution of the effect-cause relationship challenging;
- (ii) the redundancy of hundreds of thousands of InSAR observations as obtained by multi-interferogram processing of long SAR data stacks of images at high spatial and temporal resolution.

Using the Metropolitan City of Naples in southern Italy as a case study, this paper aims to discuss how a Big data analytics approach can help to handle large datasets and extract value-added information allowing the potential stakeholders to focus on the most relevant areas of concerns across the city.

2. COSMO-SKYMED BIG DATA PROCESSING

For the purposes of this research, the analysis was focused on a 12 km by 7 km wide area (of which 49 km² land), delimited by the Astroni Crater to the west and Napoli Central station to the east, and encompassing the UNESCO World Heritage historic centre of Naples and the quarters of Vomero, Antiniana, Bagnoli, Fuorigrotta and Pianura.

An unprecedented stack of 316 COSMO-SkyMed StripMap scenes acquired between 16/12/2008 and 03/08/2014 in ascending mode was processed with the MATLAB-based SAR PROcessor by periZ (SARPROZ) software based on a multi-temporal InSAR (MT-InSAR) [3]. We used a non-linear PS approach based on the assumption that the displacement trend was continuous.

PS candidates were selected based on a minimum amplitude stability index criterion and, after 2D phase unwrapping of the differential interferograms, were used to retrieve an initial estimate of height corrections, deformation, and atmospheric phase components. The network of PS was then improved by analyzing the dispersion of phase residuals and temporal coherence. All PS with amplitude stability index above 0.5 were analyzed to extract their improved height and non-linear motion time series, the latter with no a priori information on the target deformation model. 40.827° N 14.217° E was set as reference stable location (green star in Figure 1).

Short and regular revisit times of the COSMO-SkyMed acquisition plan over Naples allowed the extraction of time series with LOS displacement observations every 4-8 days on average over the 2008-2014 period, and standard errors of 1.3 mm on each observation.

411,581 PS were retrieved across the processed subset, indicating an average density of more than 8,400 PS/km² (Figure 1). In the post-processing phase, annual velocity and associated standard deviation were extracted for each PS by modelling the associated time series using a simple linear regression. The LOS velocity histogram shows a distribution centered around the value of -0.01 ± 3.46 mm/year over 5.6 years. At a glance, this value would suggest a general stability across the area of interest, although confined patterns of motions are observed (e.g. subsidence in Bagnoli and uplift in the Astroni area).

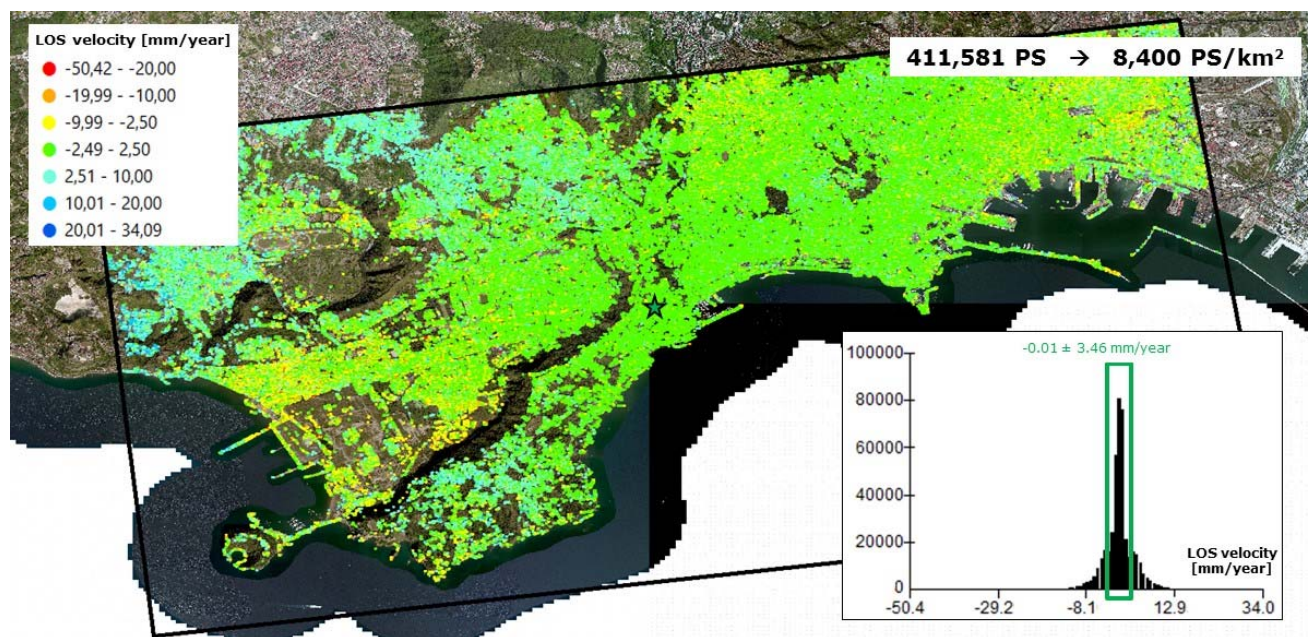


Figure 1: Line-of-sight (LOS) velocity distribution [mm/year] of ascending COSMO-SkyMed PS time series over the Metropolitan City of Naples, Southern Italy. The green star marks the location of the reference point. Product processed by Dr. P. Milillo under a license of the Italian Space Agency (ASI); Original COSMO-SkyMed product – ©ASI – 2008-2014.

3. DATA ANALYTICS APPROACH

To cream off and focus on clusters of informative PS, we applied a tiered approach of data analytics, including the following analytical steps of correlation:

- PS density vs. land cover, land use and their temporal changes as derived from EEA Urban Atlas 2006-2012;
- PS non-linear trends as per statistical analysis with the MATLAB-based PStime tool [4] vs. geotechnical zoning and lithological units defined in [5];
- PS properties (location, deformation trend, deviation from expected trend) vs. known geohazard events or mapped susceptibility.

The result of this approach was the precise identification of areas falling into one of the following categories of urban dynamics:

- Urban regeneration and new infrastructure development (e.g. Piazza Garibaldi);
- Existing built environment with coexistence of one of more factors of hazard (e.g. structural motions observed on buildings located above underground metro lines).

Furthermore, we applied PS classification indexes [6] at the scale of single building for the whole coverage of the municipal cadastral map of the city center, to achieve the finest level of hazard assessment allowed by the StripMap resolution of COSMO-SkyMed time series.

4. RESULTS AND KEY CONCLUSIONS

We run the tool PS-Time to retrieve a first order classification of the COSMO-SkyMed PS, in order to identify association between LOS displacement trends and geological properties of the city subsurface. Previous research proved that in urban environments displacement trends could be peculiar of different soil properties or informative of ground instability [7-8].

Figure 3a shows the results of the classification of the modelled time series according to the following six deformation trend types: uncorrelated, linear, bilinear, quadratic, discontinuous with constant and variable velocity. More than 50% of the PS shows bilinear trends, i.e. their time series split into two linear tracts of different velocity separated by a breakpoint in which the function is continuous, thus suggesting a trend change that may be due to a specific event. Differently, quadratic time series concentrate in the center of the processed area and highlight a general pattern of points whose velocity varies continuously in time.

Interestingly, Figure 3b highlights an association between the spatial distribution of high Annual Periodicity (AP) index (i.e. strong to very strong periodicity) and some of the geotechnical zoning units defined as per [5].

While a further dedicated geotechnical study may unlock the reason for this correlation and therefore generate value-added products, the immediate benefit of this statistical analysis is that we have cream off the redundancy of thousands of PS and narrow down to key areas of potential interest to local stakeholders.

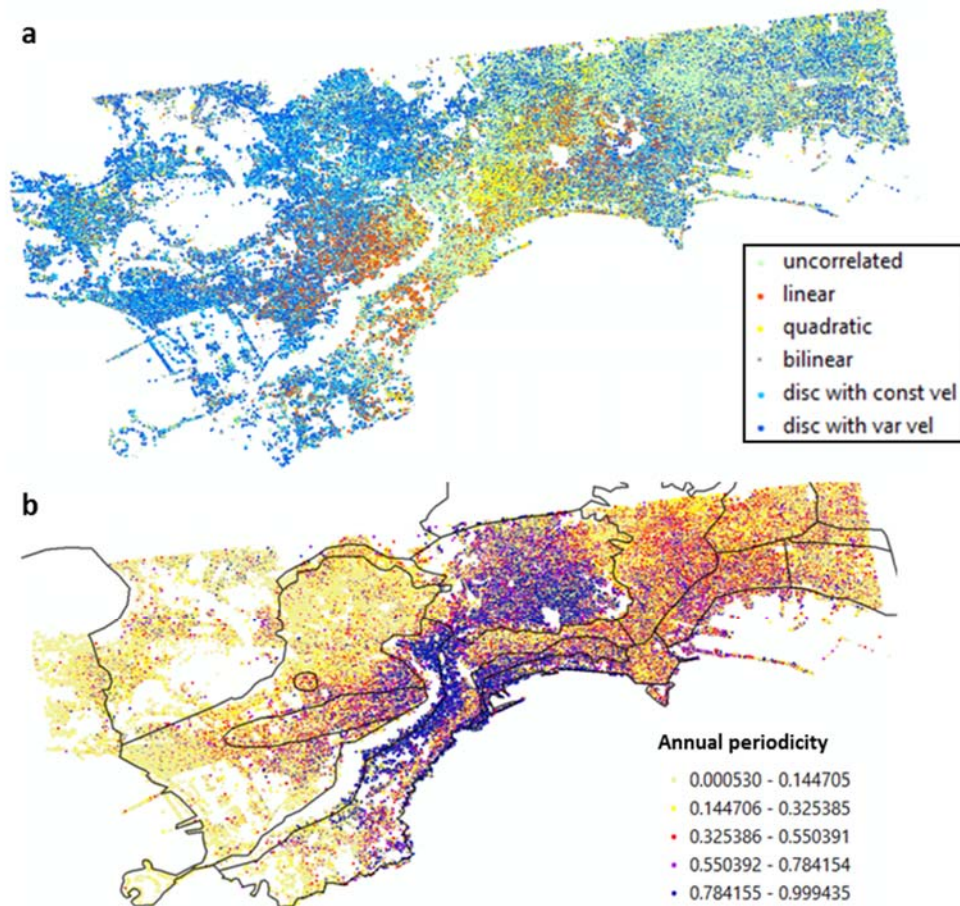


Figure 2: (a) Results of the first order classification of the COSMO-SkyMed PS by using MATLAB-based PStime tool [4] according to six different deformation trend types. (b) It is evident an association between values of annual periodicity and geotechnical zoning units [as defined in 5] in the center of the processed area. Product processed by Dr. P. Milillo under a license of the Italian Space Agency (ASI); Original COSMO-SkyMed product – ©ASI – 2008-2014.

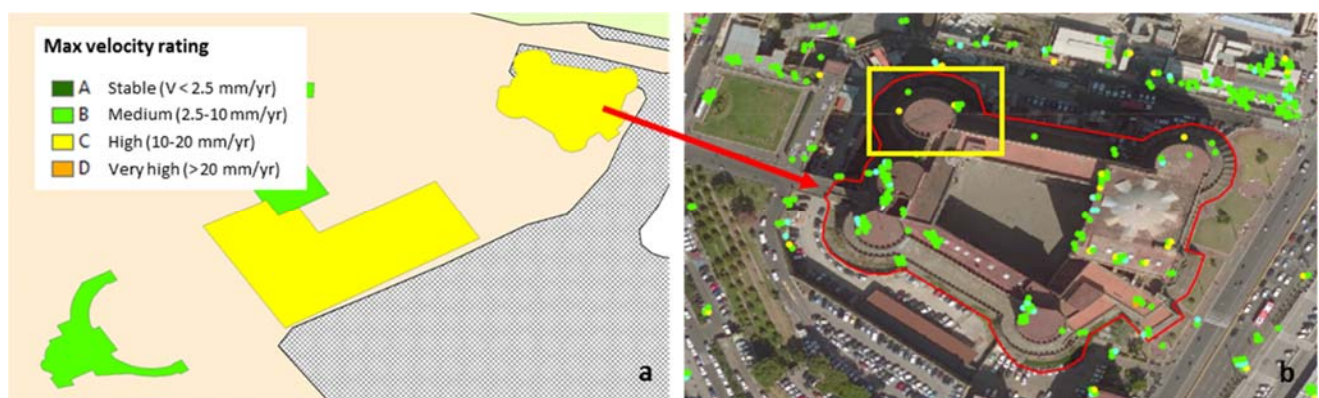


Figure 3: (a) Classification of the structural stability of the main monuments of Naples near Piazza del Plebiscito by maximum velocity rating index; (b) and zoomed view of Maschio Angioino where “high” classification correlates with unstable PS. Product processed by Dr. P. Milillo under a license of the Italian Space Agency (ASI); Original COSMO-SkyMed product – ©ASI – 2008-2014.

Similar outcome was obtained with regard to structural assessment of key assets of the city built environment by applying PS classification indexes according to the method published in [5]. Figure 3 showcases the benefit of this data analytics approach in the case of the Medieval and Renaissance castle known as “Maschio Angioino”. The “high” classification by maximum velocity rating index correlates with the presence of unstable PS in the tower facing the area of the new underground metro station.

While dedicated studies may later use this information to investigate further the structural stability of the monument (which is beyond the scope of this research), the semi-automated tiered approach developed and tested in this research proves effective to provide a granular assessment that could have been only achieved via a time-consuming manual check of the individual time series.

5. ACKNOWLEDGEMENTS

The Persistent Scatterers used in this research were provided by Dr. P. Milillo, who processed COSMO-SkyMed data accessed under ASI license agreement *UNIBAS Fenomeni Geofisici*. Project carried out using CSK® Products, © of the Italian Space Agency (ASI), delivered under a license to use by ASI.

6. REFERENCES

[1] F. Pratesi, D. Tapete, C. Del Ventisette, and S. Moretti, “Mapping interactions between geology, subsurface resource exploitation and urban development in transforming cities using InSAR Persistent Scatterers: Two decades of change in Florence, Italy,” *Applied Geography*, Elsevier, 77, pp. 20-37, 2016, doi: 10.1016/j.apgeog.2016.09.017

[2] F. Cigna, R. Lasaponara, N. Masini, P. Milillo, and D. Tapete “Persistent scatterer interferometry processing of COSMO-SkyMed StripMap HIMAGE time series to depict deformation of the historic centre of Rome, Italy,” *Remote Sensing*, MDPI, 6, pp. 12593-12618, 2014, doi: 10.3390/rs61212593

[3] Perissin D. *SARPROZ software*, Official Product Web Page: www.sarproz.com, 2015.

[4] M. Berti, A. Corsini, S. Franceschini, and J.P. Iannacone, “Automated classification of Persistent Scatterers Interferometry time series,” *Nat. Hazards Earth Syst. Sci.*, EGU Copernicus Publications, 13, pp. 1945-1958, 2013, doi:10.5194/nhess-13-1945-2013

[5] I. Alberico, M. Ramondini, and G. Zito. A database as a tool for analysis and prevention of hydrogeological instability events in an urban area: an example in the Naples area (Italy). *Italian Journal of Engineering Geology and Environment*, 2, 2006, doi: 10.4408/IJEGE.2006-02.O-05

[6] F. Pratesi, D. Tapete, G. Terenzi, C. Del Ventisette, and S. Moretti, “Rating health and stability of engineering structures via classification indexes of InSAR Persistent Scatterers,” *International Journal of Applied Earth Observation and Geoinformation*, Elsevier, 40, pp. 81-90, 2015, doi: 10.1016/j.jag.2015.04.012

[7] D. Notti, F. Calò, F. Cigna, M. Manunta, G. Herrera, M. Berti, C. Meisina, D. Tapete, and F. Zucca, “A user-oriented methodology for DInSAR time series analysis and interpretation: Landslides and subsidence case studies.” *Pure and Applied Geophysics*, Springer Basel, 172, 2015, pp. 3081-3105

[8] D. Tapete, and N. Casagli. “Application of Persistent Scatterers Interferometry time-series analysis (PS-Time) to enhance the radar interpretation of landslide movements.” *International Conference on Computational Science and Its Applications*, Springer, Berlin, Heidelberg, pp. 693-707, 2015, doi: 10.1007/978-3-642-39643-4_50

MULTITEMPORAL INTERFEROMETRY AND BIG DATA – CASE OF ALBANIA

NekiFrasheri, GudarBeqiraj, Salvatore Bushati

Academy of Sciences of Albania

ABSTRACT

The Adriatic Sea shoreline of PreAdriatic Depression in the northwestern part of Albania has experienced significant geomorphological changes during several decades with environment impact. A combination of wave erosion and subsidence is especially visible in sea transgression in Semani beaches and filling with sea waters in Patoku lagoon destroying beach resorts, and the phenomenon continues with changes visible in a span of few years with impact in human economic activities. We have experimented differential interferometry for the identification of subsidence areas using Sentinel SAR images. In order to deal with the noise due to vegetation a tentative was done to exploit multi-temporal SAR images producing summary interferograms. The work was done exploiting the cloud service offered by ESA RSS CloudToolbox service, the volume of data for multi-temporal interferograms may reach Terabytes, which is impossible to download and process locally when only modest computing resources are available.

Index Terms— ESA RSS CloudToolbox, Sentinel, SAR multitemporalinterferogram, PreAdriatic Depression

1. INTRODUCTION

During several decades particular segments of Albanian PreAdriatic Depression shoreline have experienced significant changes characterized by Adriatic Sea transgression with impact in human activities. Typical segments include Patoku Lagoon, Semani Beach, and Buna River delta (Fig. 1). A complex of factors including erosion and subsidence are considered to explain such phenomena and identify key factors for a better planning of sustainable development [1].

During the last decade we have used remote sensing to study environmental changes [2], in particular the movements of water bodies shorelines with focus on theAdriatic Sea beaches, characterized by significant sea transgression [3]. The latter phenomenon has ignited debates about the cause – erosion or subsidence. This situation led us to try the SAR differential interferometry (dinsar) for the subsidence in PreAdriatic Depression [4]. First results using Envisat images from the period 2003-2006, processed with the ESA NEST package, exposed the difficulty of using interferometry in areas covered with variable vegetation (forests, bushes, agriculture and scattered habitations), for

the southern half of the PreAdriatic Depression we received only weak signs of fringes, while for the northern part a suite of fringes correlated with the relief around Shkodra Lake and nearby Torrovisa Hills (Fig. 2). In a particular study ([5]) the Albanian Geological Service in collaboration with Italian partners identified subsidence in urban areas of southern part of PreAdriatic Depression using persistent scatterers, but the study area did not covered the Semani beach where significant sea transgression is happening.

In this work we focused on the northern part of PreAdriatic Depression where the Patoku Lagoon is situated just in south of Mati River delta. Wide sandy dunes of the lagoon were used as popular beaches and a number of buildings were constructed in seventies. Latter due to significant Adriatic Sea transgression the sand dunes were submersed and the lagoon filled with water (Fig. 1). The phenomenon is clearly visible in Landsat images. In the first differential interferogram using Envisat images (Fig. 2) fringes correlate with hilly-mountainous range that excludes the subsidence and indicates the presence of other environmental phenomena. The actual work got the hint to use multi-temporal interferograms with Sentinel SAR images from [6].

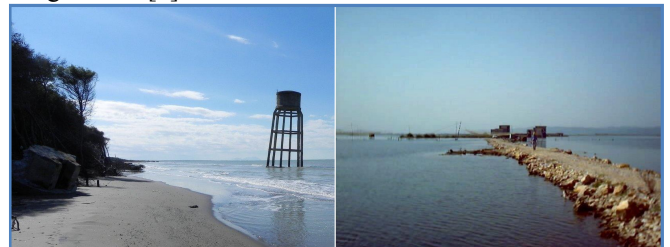


Fig. 1 – Sea transgression in Adriatic beaches: left – Patoku beach, right – Patoku Lagoon filled with sea water.

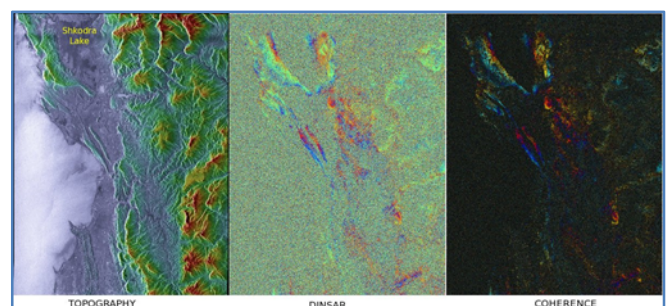


Fig. 2 – Envisat images of northern part of PreAdriatic Depression from 21Mar2003-05Nov2004.

2. USED METHODOLOGY

The used methodology consisted in producing a suite of interferograms from Sentinel-1A images for the period 2014-2017 with variable temporal steps of 12 days, 3 and 12 months. In order to reduce the random fluctuations caused by vegetation and other non-geomorphological factors, the summary of unwrapped phases was calculated for interferograms with 3 month temporal baseline, while the average was done for the wrapped phases following the idea from [6]. The line of sight displacement was calculated from the accumulative unwrapped phase both for ascending and descending satellite paths. In parallel final images of coherence, average intensity and intensity change were combined in false color RGB image, in order to better understand the weight of vegetation and urbanism in the phase noise [7].

The challenge for these calculations is related with the concept of Big Data, including tenths of GigaBytes of SAR images to be downloaded from Internet repositories and processed. It proved difficult to achieve when using relatively slow unstable Internet connections, ordinary personal computers and modest disk spaces. In this context generation of SAR interferograms may be considered as the “beginning” of Big Data from the perspective of small research units working on remote sensing.

We used computing capacities generously offered by ESA CloudToolBox [8] to speed-up the downloading of Sentinel1 images from the ESA Copernicus Open Access Hub [9]. Data processing was done using ESA SNAP toolbox [10]. Processing workflow using SNAP was the standard one [11]: TOPS Coregistration => Interferogram Formation => TOPS Deburst => Topo Phase Removal => Goldstein Phase Filtering => Snaphu unwrapping => Phase to displacement => Multilooking.

Only final products (phase, intensity, coherence, displacement) were exported as JPEG images and downloaded locally, further image processing as RGB combination and [in some cases] color enhancing was done using general purpose image processing software.

3. RESULTS AND DISCUSSIONS

The first tests were done comparing interferograms with long (two year) and short time baseline (12 days). Results were similar for different seasons, the phase and combined coherence-intensity images for the period 20 Jul 2015 – 09 Jul 2017 are given in Fig. 3. Fringes correlate with the relief.

Slopes of hills and mountains show high coherence values (red color) in the coherence-intensity image; where higher persistent values for the coherence and intensity (yellow color) are related with big cities Tirana (bottom center), Durrresi (bottom left) and Shkodra (southeastern corner of the Lake). The red spot showing high coherence with low intensity east of Tirana is related with the top of karstic Mountain with Holes.

Results of interferograms with short time baseline of 12 days are quite different from the long time baseline ones. These interferograms for different seasons were apparently similar, and the case for the period 29 Jun 2017 – 09 Jul 2017 is given in Fig. 4.

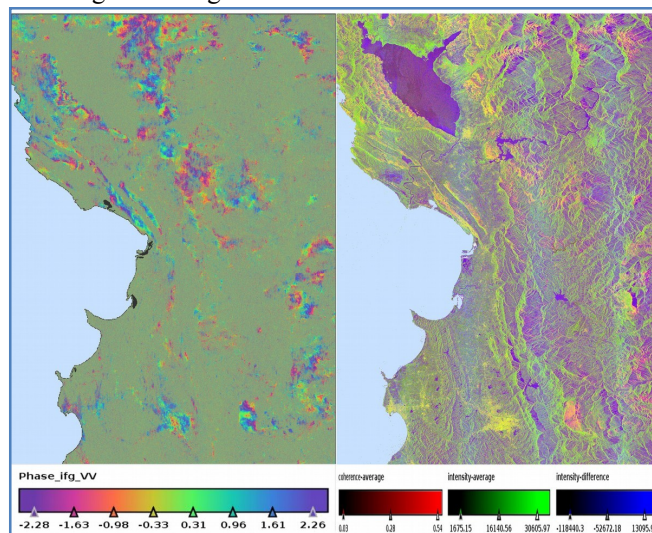


Fig. 3 – Sentinel-1A interferograms 20Jul2015-09Jul2017 (left) and related coherence-intensity image (right)

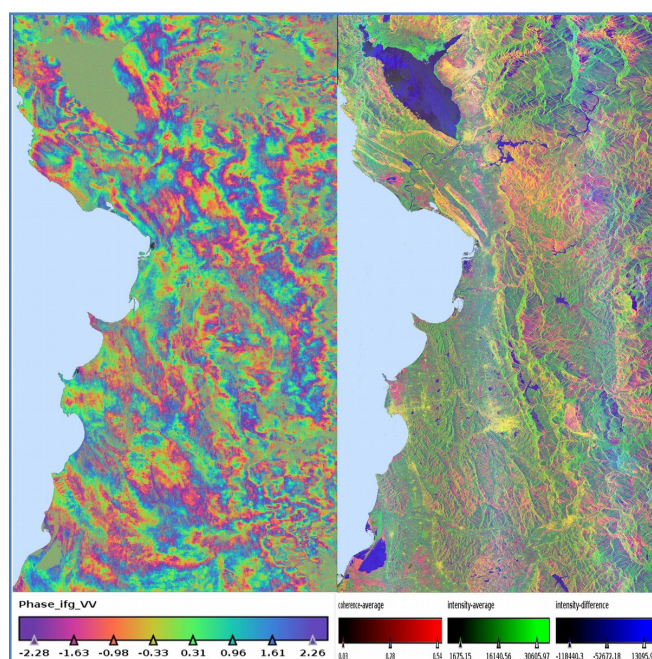


Fig. 4 – Sentinel-1A interferograms 29Jun2017-09Jul2017 (left) and related coherence-intensity image (right)

Fringes cover the whole area, having the only explanation variations in vegetation, humidity etc. In the coherence-intensity image, relative to Shkodra Lake, Torrovia hills in the south and Mirdita region in southeast are clearly visible dominated by reddish color (high coherence).

A first conclusion was that fringes correlated with high coherence were persistent in specific mountainous areas with scarce vegetation. The same can be said for big cities as well, and documented for Tirana as subsidence due to intensive construction building during last 20 years [12].

To reduce random fluctuations of the phase due to vegetation, we applied the accumulation of the phase and related line of sight displacements from 5 interferograms generated for time baseline of 3 months for the period January 2016 – April 2017. The average of the wrapped phase, coherence and intensity was calculated as well, combined in the coherence-intensity image in Fig. 5.

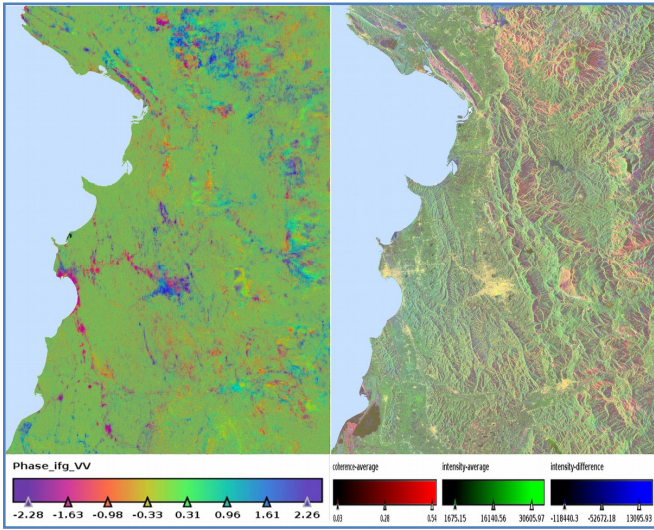


Fig. 5 – Averages of interferograms Jan2016-Apr2017 (left) and related coherence-intensity image (right)

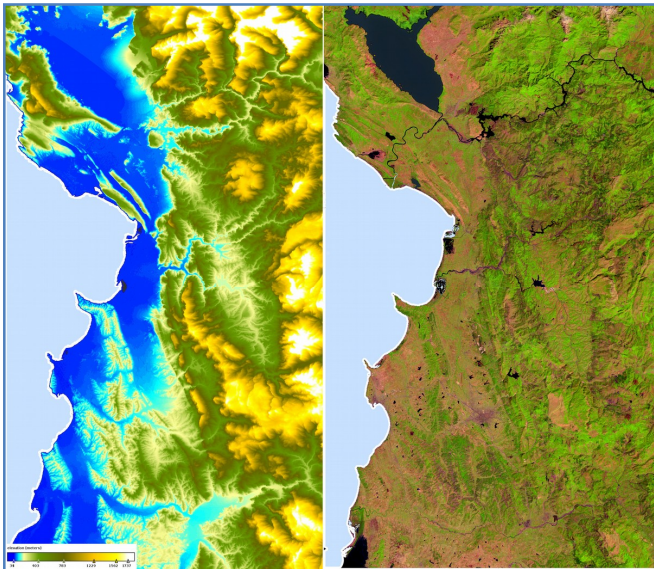


Fig. 6 – Elevation (left) and enhanced Landsat-8 image.

The average phase and coherence resulted visually improved, especially for cities of Tirana, Durresi (center-left) and Elbasani (bottom-left) where shapes or urbanized areas and the highways are visible. Scattering is persistent in

part of mountainous areas with scarce vegetation. For comparison in Fig. 6 the topography generated during the interferometry process is presented together with a view in enhanced natural colors from Landsat-8 image of 2017, 04 August [13].

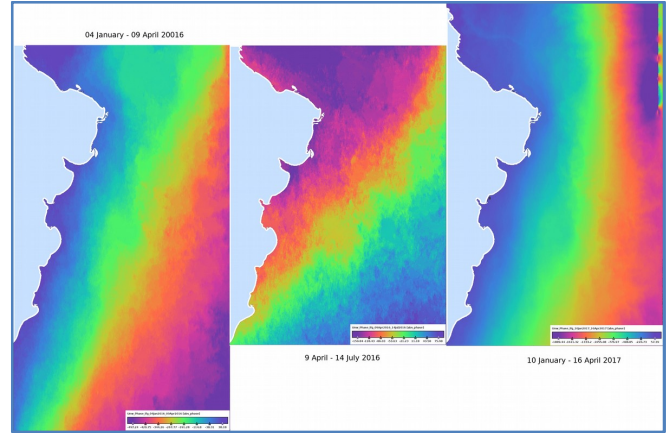


Fig. 7 – Differences between unwrapped phases from different 3 month periods

The hypothesis for observed persistent scattering and fringes are local subsidence in urbanized areas, and slow ground erosion in mountainous areas with scarce vegetation cover.

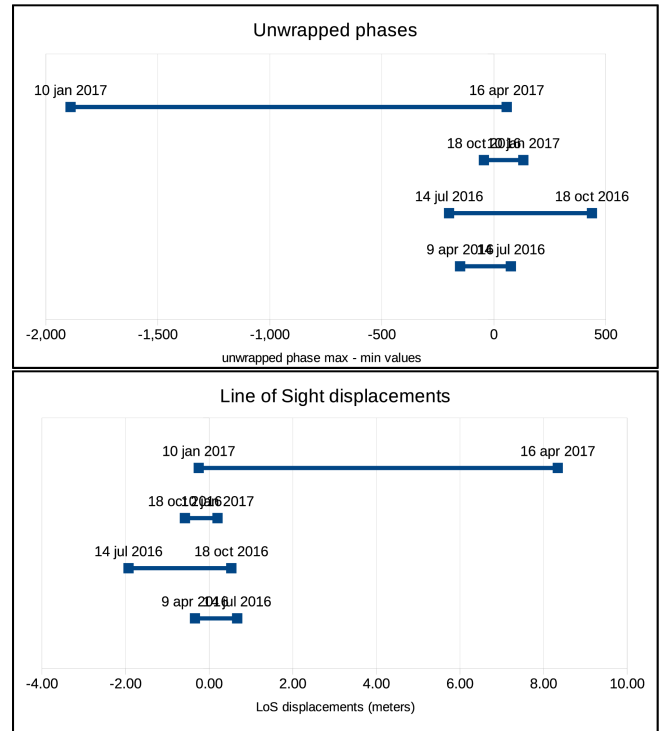


Fig. 8 – Unwrapped phase variations (top) and line of sight displacement (bottom) for April 2016 – April 2017

Unwrapped interferograms for different 3-month periods resulted not similar, as shown in Fig. 7. The variation of accumulated unwrapped phases for 3 month periods together with the variation of related line of sight

displacements is given in Fig. 8. Extreme values for periods with strong environmental changes cannot be of a geomorphological origin. Unwrapped phases and displacements from single interferograms with time baseline January 2016 – January 2017 are given in figures Fig. 9 (ascending orbit) and Fig. 10 (descending orbit).

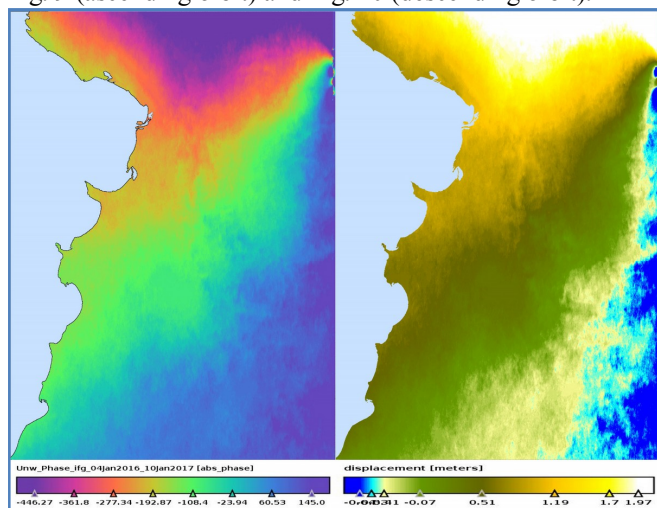


Fig. 9 – Unwrapped phase (left), line of sight displacement (right) for ascending orbits 04 Jan 2016 – 10 Jan 2017

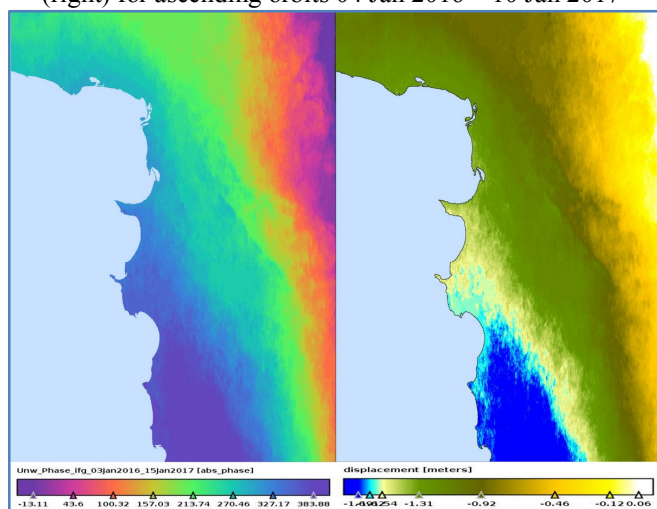


Fig. 10 – Unwrapped phase (left), line of sight displacement (right) for descending orbits 03 Jan 2016 – 15 Jan 2017

4. CONCLUSIONS

Multitemporal interferograms from Sentinel SAR images show slow erosion due to atmospheric factors in mountain ranges, and subsidence in main cities due to weight of new buildings. Variations of phases, orientation of anomalies, extreme values of phases and displacements indicate the complexity of factors that impact the SAR data in territory of Albania.

It seems not possible to investigate the existence of subsidence in PreAdriatic Depression through simple application of interferometry workflows, and that methods based on multitemporal persistent scatterers are necessary.

5. REFERENCES

- [1] S. Aliaj, G. Baldassare, and D. Shkupi, "Quaternary subsidence zones in Albania: some case studies". Bull EngGeolEnv 59:313-318 Springer Verlag 2001.
- [2] N. Frasheri, G. Beqiraj, S. Bushati, and A. Frasheri, "Application of Remote Sensing for the Analysis of Environmental Changes in Albania". ESA Living Planet Symposium, 9-13 May 2016, Prague Czech Republic.
- [3] N. Frasheri, G. Beqiraj, S. Bushati, A. Frasheri, and E. Taushani, "Remote Sensing Analysis of the Adriatic Shoreline Movements". Conference on Integrated Coastal Zone Management in the Adriatic Sea, Kotor, Montenegro 29 Sept – 1 Oct 2014.
- [4] N. Frasheri, G. Beqiraj, and S. Bushati, "Investigation of environmental changes using satellite radar images – experiences from Albanian hot spots". EO Open Science Conference 2016, Frascati 12-14 September 2016
- [5] M. Lamaj, N. Frasheri, S. Bushati, L. Moisiu, G. Beqiraj, and A. Avxhi, "Application of Differential Interferometry for Analysis of Ground Movements in Albania". Fringe 2015 Workshop 23–27 March 2015 Frascati, Italy
- [6] X. Xu, D.T. Sandwell, E. Tymofyeyeva, A. Gonzalez-Ortega, and X. Tong, "Tectonic and Anthropogenic Deformation at the Cerro Prieto Geothermal Step-over Revealed by Sentinel-1 InSAR," ESA Fringe 2017 Workshop, Helsinki 5-9 June 2017.
- [7] U. Wegmüller, M. Santoro, C. Werner, and O. Cartus, "On the Estimation and Interpretation of Sentinel-1 Tops InSAR Coherence," 9th International Workshop Fringe 2015, Frascati, Italy 23–27 March 2015.
- [8] ESA, CloudToolBox, <http://eogrid.esrin.esa.int/cloudtoolbox/>
- [9] ESA, CopernicusOpenAccessHub, <https://scihub.copernicus.eu/>
- [10] ESA, Sentinel Application Platform (SNAP), <http://step.esa.int/main/toolboxes/snap/>
- [11] ESA, Sentinel-1 Toolbox Tutorials, <http://step.esa.int/main/doc/tutorials/sentinel-1-toolbox-tutorials/>
- [12] S. Kuçaj, Use of Interferometry Method SAR for Evaluation of Subsidence of Engineering Buildings with Application Example from City of Tirana. PhD Thesis, 2016 (in Albanian). <http://dibmin-fgjm.org/doktorata/DisertacioniSKucaj.pdf>
- [13] USGS, Earth Explorer, <https://earthexplorer.usgs.gov/>

CLIMATE EXTREME EVENTS DETECTION BASED ON WEATHER FORECASTING VARIABLES COMBINATION

Javier Lozano Silva

Mondragon University
Intelligent Systems for Industrial
Systems Research Group
Arrasate-Mondragon

Marco Quartulli, Igor G. Olaiozola

Vicomtech-IK4
Data Intelligence for Energy &
Industrial Processes
Donostia-SanSebastian

ABSTRACT

Nowadays the climate change is one of the powerfuller challenges that the humanity can deal. The causes and effects produced by it are more frequently and most catastrophics, and damages could be material or human. Trying to minimize these tragically effects, this work tries to deal the prediction of extreme meteorological events based on the analysis of meteorological forecasting data and applying machine learning algorithms. Obtained results are presented as thematic map representing risk percentage.

Index Terms— Meteorological forecasting, climate extreme events

1. INTRODUCTION

The consequences of extreme weather events differ based on the location of the event, the terrain in which they happens or the affected infrastructures.

Meteorological predictions can be made in different spatio-temporal contexts, from very short to long time spans and from local to global in spatial domains. Nowcasting methodologies try to address the limitations of global predictions over long time spans by local predictions over very short times.

The limited availability of expert meteorologists represents an opportunity for the development of efficient Machine Learning methodologies to support in particular the detection of extreme events in forecasted data cubes.

2. STATE OF THE ART

Extreme weather forecasting is the subject of a number of papers [1, 2, 3, 4, 5].

The work in [1] is only centered in heavy rainfall prediction. They proposed a model based on deep neural network

Javier Lozano Silva was part of the Data Intelligence for Energy & Industrial Processes research team in Vicomtech-IK4 before this submission.

from previous climatic parameters. The model is able to predict extreme events from 6 to 48 hours before they occurs. Obtained results are better than other methods as the time to occur increases.

In [2] combines an non lineal attribute selection method with a regression model based on support vector machine algorithm to obtain rainfall prediction. Obtained results shown that proposed model has better Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) that other works in the literature. The data used in this work covered the period between 2005 and 2014 from monitoring stations from Taiwanese mountains.

Based on the combination of different meteorological parameters like humidity, temperature, pressure, wind velocity and others, a system for weather prediction, based on the combination of a neural network and fuzzy logic system, is presented in [3]. It is also able to verify how one weather parameter affect another.

Beyond the pure weather prediction, can be found different approaches in the literature that shown the use of prediction data in different contexts. The use of the weather prediction to inform drivers is target of the work presented in [4]. An autonomous system for short time weather conditions and system for derived risky situations in roads are presented.

In [5] short time fog forecasting system is presented. Based on BayesNet algorithm is able to predict for next 1, 2 and 3 hours. Applied in Charles de Gaulle airport in Paris use the data available by SYNOP system from the last 17 years.

Contributions such as [6] show a combination of machine learning algorithms with map servers, via web interface, to allow the dynamic creation of a thematic maps based on user selected patterns.

3. METHODOLOGY

Different contexts and users require ways to analyse meteorologic data with the objective to create thematic maps adapted to their priorities. The main idea of the present contribution is the combination of short time global forecasting techniques

with Machine Learning algorithms to obtain detectors for extreme forecasted events that exhibit high precisions and high specificities, being adapted to the needs and characteristics of a specific application domain (e.g. wineries, wind power generation), thereby creating thematic meteorological maps as per Figure 6.

The solution adopted is based on "Connected Scatter Plots" such as the ones in [7]. This kind of visualization combines variables in a series of two dimensional Cartesian planes. The linking of different points by lines creates visual patterns that can be the subject of analysis. Meteorological extreme events will be displayed as plots with large covered areas. Figure 1 presents a few examples of this visualization mode. Secondary feature extraction methodologies inspired by Computer Vision techniques can be used to transform this primary feature space describing time series of meteorological forecasting data to numeric morphological features extracted from the created figures.

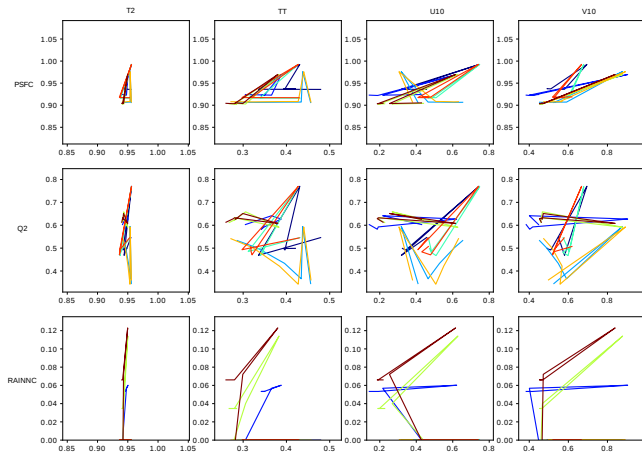


Fig. 1: Connected Scatter Plot primary feature space

Possible secondary descriptors include for instance the relation between the convex hull of the time series points and the square, for dimensionless relation, of the longitude of the path. In Figure 2, these new secondary feature extraction process is presented, and in the Figure 3, the relation between these new complex variables are presented. We observe that in this figure we can start to detect extreme event like clusters of points grouped around top right corner.

Available forecasting data is reduce geographically to Iberian Peninsula area and the next 48 hours for each day. In order to carry out the execution of the prediction model we have used the model Weather Research and Forecasting, WRF. It is a limited area model, non-hydrostatic and mesoscale, widely diffused in research fields like Meteorology and Atmospheric Sciences. From this model, obtained each forecast data set is composed by 34 bidimensional variables, most of them with 25 elevation levels, with a size of 390 x 441 pixels. It all adds up to near 20GB of data daily.

Each instance of data cubes is formed by previous 6 hours

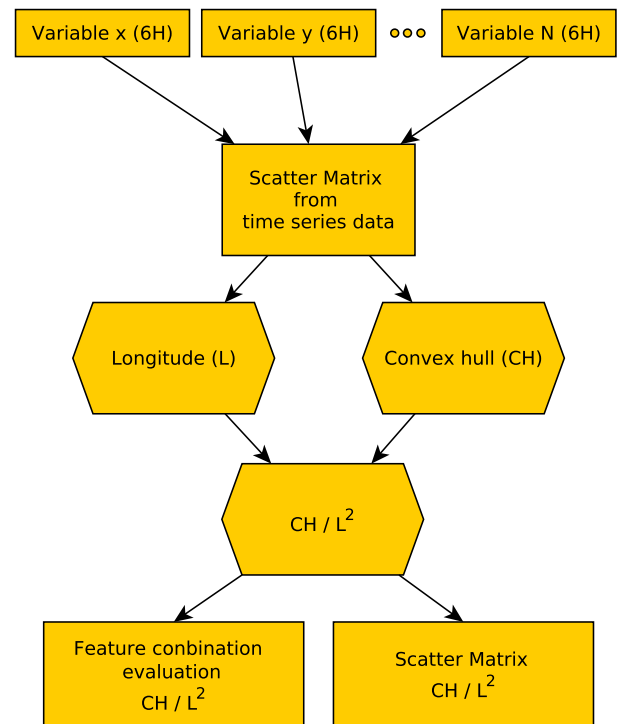


Fig. 2: Secondary feature extraction process

of the target hour and are created by the combination of the following descriptors: PSFC (pressure in the surface), T2 (temperature at 2 meters over sea level), Q2 (water vapor ratio at 2 meters over sea level), RAINNC (total accumulated precipitation) and wind definition variables at 10 meters over sea level U10, x component and V10, y component.

The different characteristic of the climatic areas existing in the Iberian Peninsula force us to analyzed them in a separated way. The areas are named as C026, C027, C031, C032, C034 and C035.

Once the values of secondary descriptor variables are obtained, Machine Learning processes is applied [8]. To evaluate the performance of the system, standard measures including Precision, Recall, F1 and Specificity can be used. We consider and compare different algorithms including SVM, Neural Networks and Adaboost. At the same time, different Cross Validation [9] configurations are used to compare the performance and standard measures.

4. RESULTS

Following the recommendation in [10], obtained secondary descriptors are analyzed making a range of importance of them.

Obtained results by different SVM and Neural Network algorithm with different configurations weren't significant. The result obtained with Adaboost algorithm [11], configured

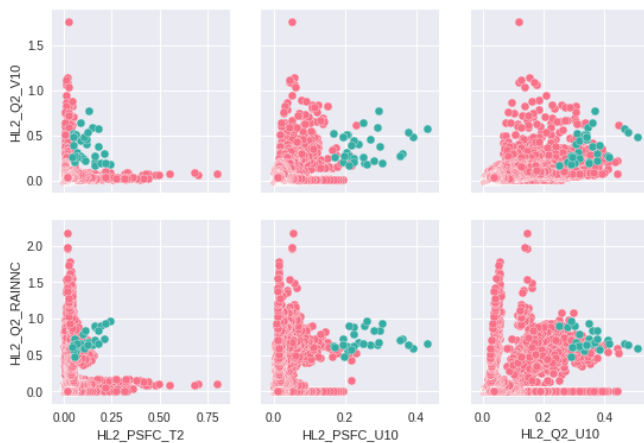


Fig. 3: Secondary feature space example

with DecisionTreeClassifier, are shown in Figures 4 and 5. In first place the results obtained with the best three secondary descriptors are shown and then the results obtained using all secondary descriptors.

As can be seen in these figures the use of all secondary descriptors has different effects. In one hand these allow to obtain better results in different climatic areas, especially in two of them, named as C031 and C034, but in other hand the results in climate C035 are worst. On the other hand, the use of different test percentage shows that has not an excessive improvement inside each climatic area. Considering the grade of unbalanced in the data, in the sense of extreme events, the Adaboost algorithm shown a great performance.

Finally in Figure 6 an implementation of the prototype is presented. The analysis of the meteorological data cubes from the Iberian Peninsula area becomes into a thematic map for a concrete day and hour (2016-02-14 11H). In orange detected extreme events in the Northeast of Spain.

5. CONCLUSIONS

The presented treatment indicates that automatic methodologies based on Machine Learning can efficiently support the interpretation of large scale local and short range meteorological forecasting results, in particular for the detection and characterization of extreme events in the forecasted data cubes, based on a small set of examples put forward by expert supervisors.

6. REFERENCES

- [1] Sulagna Gope, Sudeshna Sarkar, Pabitra Mitra, and Subimal Ghosh, *Early Prediction of Extreme Rainfall Events: A Deep Learning Approach*, pp. 154–167, Springer International Publishing, Cham, 2016.
- [2] Jun-He Yang and Ching-Hsue Cheng, “A novel rainfall forecast model based on integrated non-linear attributes selection method and support vector regression,” in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Aug 2015, pp. 989–994.
- [3] Sanjay Khajure and S.W. Mohod, “Future weather forecasting using soft computing techniques,” *Procedia Computer Science*, vol. 78, pp. 402 – 407, 2016, 1st International Conference on Information Security and Privacy 2015.
- [4] V. R. Tomás, M. Pla-Castells, J. J. Martínez, and J. Martínez, “Forecasting adverse weather situations in the road network,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2334–2343, Aug 2016.
- [5] G. Zazzaro, G. Romano, P. Mercogliano, V. Rillo, and S. Kauczok, “Short range fog forecasting by applying data mining techniques: Three different temporal resolution models for fog nowcasting on cdg airport,” in *2015 IEEE Metrology for Aerospace (MetroAeroSpace)*, June 2015, pp. 448–453.
- [6] J. Lozano Silva, N. Aginako Bengoa, M. Quartulli, I.G. Olaizola, and E. Zulueta, “Web-based supervised thematic mapping,” *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. PP, no. 99, pp. 1–12, 2015.
- [7] T. N. Dang, A. Anand, and L. Wilkinson, “Timeseer: Scagnostics for high-dimensional time series,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 470–483, March 2013.
- [8] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth, “Crisp-dm 1.0 step-by-step data mining guide,” Tech. Rep., The CRISP-DM consortium, August 2000.
- [9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- [10] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts, “Understanding variable importances in forests of randomized trees,” in *Advances in Neural*

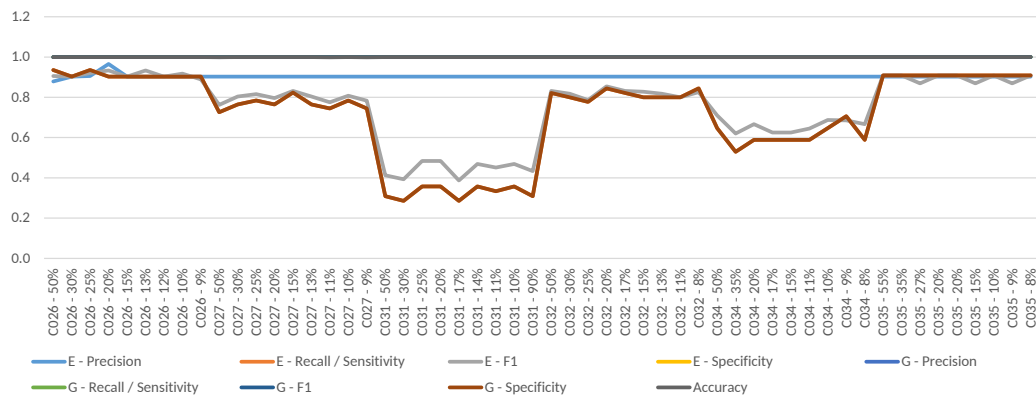


Fig. 4: Adaboost statistical results with best three secondary descriptor variables (PSFC-T2, Q2-PSFC and PSFC-U10). In X axis the climatic area name and used test percentage for cross validation.

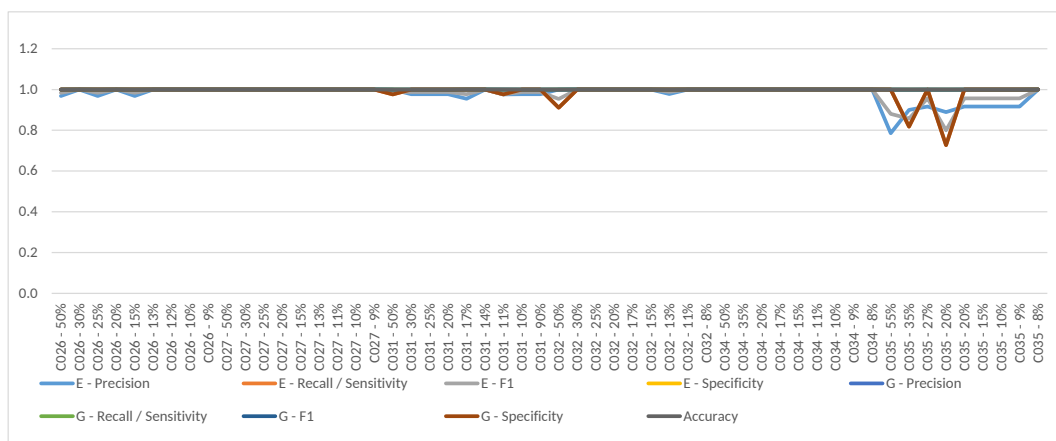


Fig. 5: Adaboost statistical results with all second descriptor variables

Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., pp. 431–439. Curran Associates, Inc., 2013.

- [11] Yoav Freund and Robert E Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, 1997.

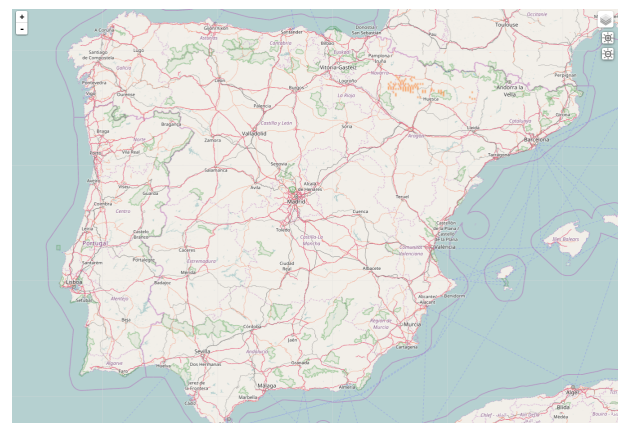


Fig. 6: Representation of extreme event prediction in orange in the Northeast of Spain

MEDUSA: MULTITEMPORAL EARTH OBSERVATION DATAMASS FOR URBAN SPRAWL AFTERCARE

Elise Koeniguer, Karine Adeline, Jérôme Besombes, Alexandre Boulch, Xavier Ceamanos, Adrien Chan Hon Tong, Guillaume Dufour, Fabrice Janez, Aurélie Michel, Aurélien Plyer, François Rogier, Pauline Trouvé-Peloux

Onera – Chemin de la Hunière et des Joncherettes, 91123 PALAISEAU Cedex, France

ABSTRACT

Medusa project is designed to bring together and to promote processing of remote sensing images in the current context of big data, with a focus on developing a demonstrator, for four different scenarios: updating of the 3D GIS database, traffic and parking monitoring, heat islands, and building deformation monitoring. Important scientific innovations have been made, such as rapid coregistration, the use of deep learning, fusion of local sensor data with remote sensing and colored composite for the visualization of radar temporal series.

Index Terms— Remote sensing, big data, smart city, temporal series

1. INTRODUCTION

The new context of big data in Earth Observation is illustrated by the launch between 2014 and 2020 of a network of satellites called Sentinel, which already acquire images that are immediately distributed worldwide as open data. Meanwhile, data hubs provide free and open access to rolling repositories of Sentinel products, while exploitation platforms are also developed to allow image processing directly in virtual workspaces with the general platform capabilities. This new context of big data for Earth observation, particularly in the field of urban development, has led Onera to launch a new research project, called "Medusa". The acronym stands for "Multidate Earth observation Data for Urban Sprawl Aftercare". It includes the development of new application for sustainable management of city, or smart cities [1].

Medusa project is designed to bring together and to promote processing of remote sensing images in the current context of big data, with a focus on developing a demonstrator. This prototype is intended to show how temporal data stacks obtained from a variety of sources are an opportunity for new application frameworks. The main idea is to start from an existing 3D GIS database that best describes the scene, and to enrich the pool of information that is necessary for enlightened policy decisions. As soon as a new image is acquired, we try to answer the question: what is the best way to improve the product for our application using this remote

sensing image. Four different applications are considered in this project, as illustrated in Fig.1:

1. Monitoring the development of the urban extension, both by a 2D (ground cover) and 3D (volume representation) update mapping
2. Monitoring traffic and parking occupancy
3. Mapping Urban Heat Islands.
4. Monitoring ground subsidence and structure deformations.

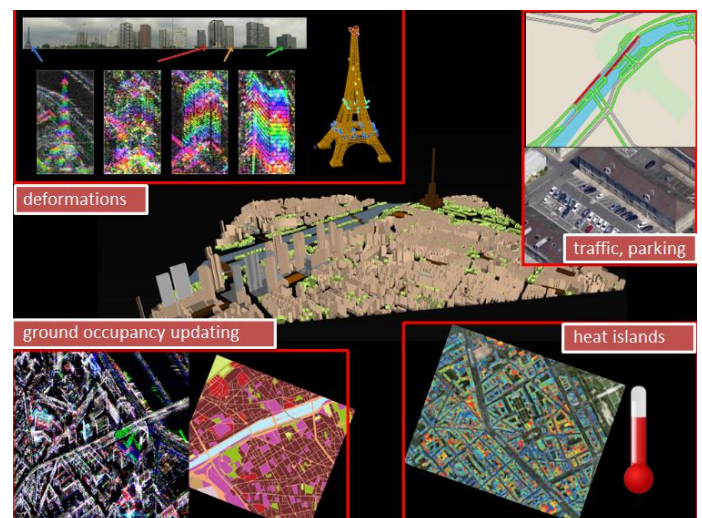


Fig. 1. Four different scenarios using Earth Observation for urban monitoring

Thus, the project aims at meeting the double challenge of the huge volume of data and its inhomogeneity. In many cases, generic treatments will be required, such as coregistration [2], change detection, filtering methods, specific correction and calibration. New methods of machine learning on large data sets, such as recent deep learning methods, are also explored [3]. Finally, specific algorithms will be developed for each of our four applications. They are now addressed in the following sections.

2. CHANGE DETECTION

For the monitoring of urban expansion, the main idea is to start from an already existing topographic base, and to update it as often as possible. First, 2D change detection is

envisaged, preferably in radar, because it is the sensor which offers the best change detection performance: being an active sensor, the signal is stable temporally. Characterization of changes can be driven using optical data or also using a search for images published on the unrestricted public Internet. Another need concerns the classification of the new zone that has undergone a change; It will be considered by deep learning classification methods.

The axis of scientific innovation concerns the fact that the change detection is extended to a temporal dimension; it is no longer restricted to the comparison of two images, but can be considered as a method of temporal analysis. Today, it is possible to benefit from large time series of data; whether commercial or in open source. In this framework, the aim of the research work is to enrich our theoretical and practical knowledge on the SAR phenomenology of the temporal profiles.

In this context, the notion of change is divided into several different scenarios: large-scale detection of areas of human activity, and capacity for distancing between changes on natural areas (fields, mountain) and on human targets.

Our first work deals with the temporal analysis of the profiles by distinguishing between:

- stable natural areas that corresponds to developed speckle and permanent scatterers.
- permanent changes and one-time changes

One of the first results consisted in the development of two close RGB composition maps highlighting these multi-temporal notions. The first one is dedicated to very high resolution images such as TSX images; second one is dedicated to Sentinel stacks.

In both cases, the colored representation is coded in the HSV space. The hue is given by a color assigned to the date for which the maximum intensity is obtained. The saturation is established on the criterion of temporal stability. It is the ratio of the standard deviation to the temporal average. This ratio is constant for a Rayleigh distribution, whatever the intensity. It is low on permanent scatterers: in this case of very stable target, no color will be privileged. On the contrary, it is particularly high as soon as one exits from these two cases, that is to say, when we have a notable event in the time profile. Finally, the choice of intensity will

depend on the resolutions and the type of change that one wishes to put forward

For Sentinel images, the visual gain from spatial averaging is particularly interesting. Also, the colored intensity will be coded by the time average of the intensities. On these images, the fields and new buildings appear in color.

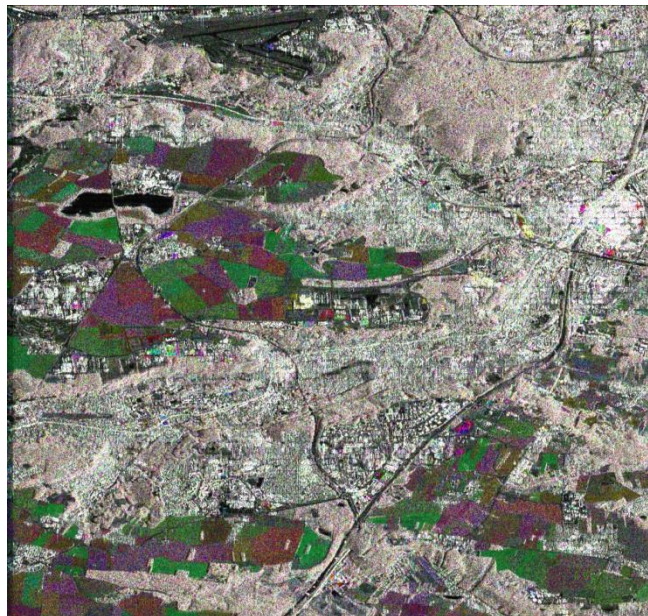


Fig. 2. Colored composition of 65 Sentinel 1 images acquired over Palaiseau/Saclay (Descending). Saturation: ratio standard deviation/average, Intensity: temporal-average, Hue: date of maximal intensity.

For TerraSAR-X images, the resolution is fine enough to capture all the changes related to one-time events: for example, the presence or not of a boat. In this case, in order to ensure visualization of these targets in the image, it is necessary to choose the maximal intensity values of the profile for the intensity; otherwise the average value will not allow visualizing this type of target. In this composition on Fig 3, the colors make it possible to clearly identify several construction areas, as well as the flow of barges synthesized with a range offset due to their doppler for those that were in motion.

Future works will concern the development of the detection /classification of the different types of change and to extend it on a large scale on the Sentinel data.

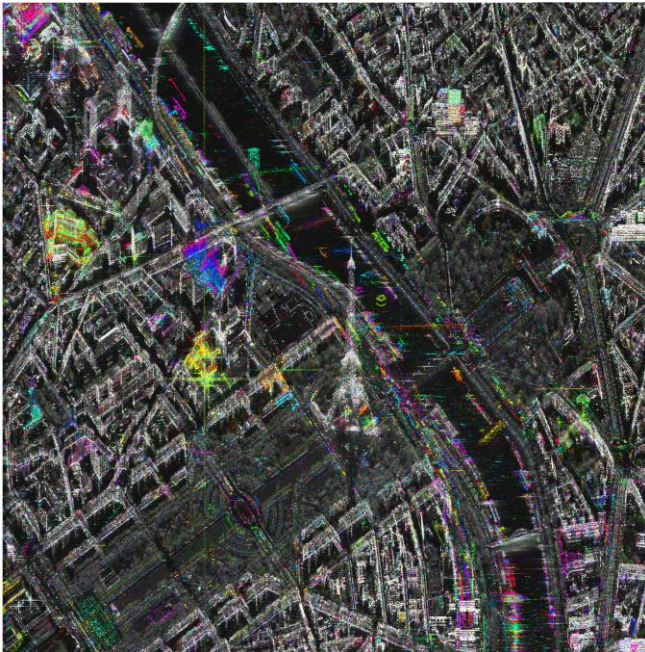


Fig. 3. Colored composition of 98 TerraSAR-X images acquired over Paris. Saturation: standard deviation/average ratio, Intensity: maximal intensity, Hue: date of maximal intensity.
(Temporal Stack from DLR Project LAN2939)

3. TRAFFIC MONITORING

Traffic monitoring will be based on detection algorithms, designed for radar or optical sensors. The project investigates to what extent a good, even if not perfect, vehicle detection map could serve as an initialization procedure for traffic flow models. The advantage is to have this initialization on a large area. Detection map fused with the road network maps will be converted into traffic density maps. It will initialize the classical models of traffic that apply the rules of fluid dynamics to traffic flow. Next, inputs of the model will be measures obtained by the traditional in-situ technologies such as piezoelectric sensors, magnetic loops or video sensors..

The demonstration of the contribution of such an initialization map can be made using local data and two airborne remote sensing acquisitions at a few hour intervals with the following steps:

1. Vehicle detection and traffic density conversion to serve model initialization
2. Application of the flow model to predict a later state, using the local data.
3. use of the new remote sensing image as a ground truth, to verify the prediction of the model

Initial efforts are focused on vehicle detection on high-resolution TSX or optical data. Recently, nice first results (Fig. 4) were obtained through the application of SegNet

[4], a deep network algorithm, for the detection of vehicles on an aerial photography, based on a training performed over more than 20000 vehicles.

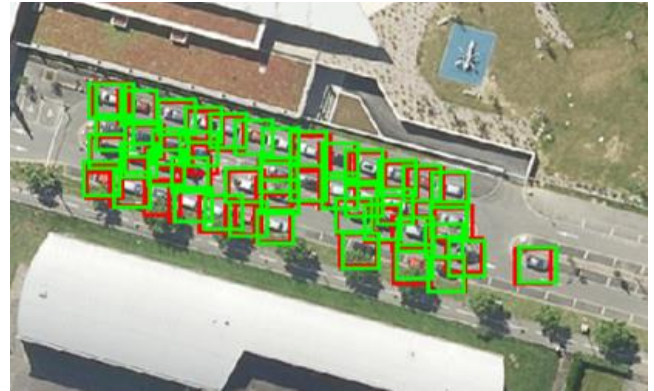


Fig. 4 Results of car detection (Top) on a parking and (Bottom) on a road using SegNet [4]
Red: Ground truth. – Green: detection

4. URBAN HEAT ISLANDS

For the monitoring of urban temperatures, the approach chosen in the project is the combination of infrared thermography and multispectral/hyperspectral images in the thermal infrared and reflective domains. It is known that large urban areas often experience higher temperatures, greater pollution, and more negative health impacts during hot summer months. These effects are the result of building material that can be heat-absorptive, such as dark pavement and roofs, heat-generating activities (such as engines and generators), and the absence of vegetation (which provides evaporative cooling). Thus, hyperspectral images make it possible to establish a classification map of the materials present, while the thermography allows a temperature mapping.

Current efforts include the organization of a hyperspectral and infrared airborne measurement campaign on the city of Toulouse. At the moment, data processing algorithms are designed and evaluated on data of the satellite type at less fine resolutions, in particular through the CATUT project [5].

5. LAND SUBSIDENCE AND STRUCTURE DEFORMATION MONITORING

For this scenario, the idea is to see to what extent remote sensing can monitor the deformations, both at the level of the ground but also of the structures. SAR technology is being considered for this purpose. The measure of deformation of the ground is already well known by the DInSAR technics; it can be envisaged Sentinel 1 data at C-band. The measurement of structural deformations is more exploratory and requires merging the interferometric repeat pass data with measurements of fine elevation. The approach proposed here is based on the combination of TerraSAR-X and a 3D-data base. In this case, a method has been proposed to follow the deformations of a resolved structure. It is fully described in [6]. The originality of the proposed method is multiple:

- No hypothesis is made on the type of deformation. Thus chaotic deformations can also be analyzed.
- The estimation of the deformations is done without a priori selection of permanent scatterer, but is performed for all the pixels belonging to the building faces seen in the image.
- A specific phase unwrapping algorithm is proposed to unwrap the phases along the range axis. It is, based on the use of circular moments, which does not require any spatial average of the phase. The entire spatial resolution of the information is thus preserved.
- Finally, the contribution of the polarimetry was analyzed (at reduced resolutions in this mode): it was shown that the polarimetric information makes it possible to get fine information on the orientation of the facade elements and their structural complexity.

6. AN OPEN-SOURCE TOOL OF COREGISTRATION GEFOLKI

For this project, developed codes will be distributed under the GNU GPL License. To date, a coregistration tool for remote sensing images, GeFolki, is already available. The aim of this tool is to calculate the fine residual deformations between two remote sensing images that can still exist, even after georeferencing.

The most promising frameworks of use that have already demonstrated their effectiveness are the following:

- the computation of interferograms, under a priori difficult conditions (strong temporal decorrelation, different acquisition modes or different resolutions: stripmap/spotlight [2], etc.), or for large image stacks (Sentinel 1)
- the fine superimposition of heterogeneous images (optical / Lidar, radar / Lidar, radar / optical) where georeferencing is not precise enough. For example, the fine coregistration of a LIDAR DEM image with a hyperspectral image may be particularly useful for improving atmospheric corrections algorithms used for the calibration of hyperspectral images [7].
- the fine coregistration of images acquired by the same airborne platform using two different sensors with different distortions and focal lengths (SWIR / VNIR for airborne hyperspectral data.)

7. CONCLUSION

The MEDUSA project develops innovative works to use remote sensing as an urban surveillance tool; four different scenarios are under development. The most promising of these will be used to demonstrate feasibility on a global scale. The progress of the project can be consulted on the site w3.onera.fr/medusa

The TSX images used in this paper come from the DLR which we thank warmly.

8. REFERENCES

- [1] Baghdadi N. & Zribi, M. (2016). Land Surface Remote Sensing in Urban and Coastal Areas, Chapter Optical Remote Sensing in urban environment (X. Briottet, C. Weber, R. Oltra-Carrio, N. Chehata, A. Le Bris), ISTE Press, London and Elsevier
- [2] Plyer, A., & al. (2015). A New Coregistration Algorithm for Recent Applications on Urban SAR Images. *Geoscience and Remote Sensing Letrs, IEEE*, 12(11), 2198-2202.
- [3] Audebert, N., Le Saux B. & Lefèvre, S. (2016) How useful is region-based classification of remote sensing images in a deep learning framework?, *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, pp. 5091-5094.
- [4] Badrinarayanan, V., Kendall A., & Cipolla, R. SegNet (2017): A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.PAMI, 2017
- [5] Adeline, K., & al. (2017) Challenges of the franco-indian multispectral thermal spatial mission for urban heat islands monitoring, 5th International Symposium Recent Advances in Quantitative Remote Sensing (RAQRS), 18-22th September 2017, Torrent (Valencia), Spain,
- [6] Weissgerber, F., Colin-Koeniguer, E., Nicolas, J. M., & Trouvé, N. (2017). 3D Monitoring of Buildings Using TerraSAR-X InSAR, DInSAR and PolSAR Capacities. *Remote Sensing*, 9(10), 1010.
- [7] Ceamanos, X., Briottet, X., Roussel, G., & Gilardy, H. (2016). ICARE-HS: atmospheric correction of airborne hyperspectral urban images using 3D information. In *SPIE Remote International Society for Optics and Photonics*, October 2016

BIG EARTH OBSERVATION DATA FOR FAST DETECTION OF DEFORESTATION USING ADAPTATIVE FILTERING

Alber Sánchez, Gilberto Câmara

National Institute for Space Research
Earth Systems Research Center
São José dos Campos, SP, Brazil

ABSTRACT

Nowadays, several terabytes of remote sensing data are available for researchers to track changes on Earth's surface. These data can help scientists to make a significant contribution in the early detection of deforestation. On the other hand, engineers have been developing adaptive filtering methods for control-related applications because of their reliability and low computation cost. In order to timely detect deforestation events, we introduce adaptive filtering as a light weight alternative for processing massive sets of time series of tropical forest data. Our preliminary results show that, methods such as the Kalman filter are suitable for early detection of deforestation in the Amazon forest while keeping a balance between speed and accuracy.

Index Terms— Earth observation, time series, adaptive filtering, array databases

1. INTRODUCTION

Forest loss is disturbing because it affects the quality of water and air, the carbon cycle, and the preservation of biodiversity. Despite its importance, the available global accounts of deforestation are scarce and subject to semantic and local inaccuracies. Besides, these accounts are useless for deforestation prevention because of the amount of time required to process large amounts of imagery. Unlike these accounts, forest monitoring programs target specific forests aiming to prevent and discourage deforestation [1, 2, 3].

However, the existing forest monitoring programs could be reaching their limits. Take for example successful projects such as Brazilian PRODES (Program for Deforestation Assessment in the Brazilian Legal Amazon using remote sensing images and digital image processing) and DETER (Project for real-time detection of deforestation). They rely on satellite imagery, hardware, software, and human criteria for producing high quality deforestation assessments which played a key role in deforestation reduction during the first decade of

the 21st century. But the permanent improvement on imagery resolution demands additional human and computational resources. PRODES and DETER analyze the *Legal Brazilian Amazon*, which is 60% of the whole Amazon forest, and they detect deforestation patches larger than 6.25 ha. Both projects are exclusively focused on intact forest for which they use a mixture of unsupervised and supervised classification methods. These methods are unable to keep pace with the increasing volume of data due to finer spatiotemporal resolution of imagery and emerging challenges such as forest degradation, conversion, or modification. Finally, their effectiveness is undermined by deforesters who are learning how to avoid detection [4, 5, 6].

In the Earth Observation (EO) field, the traditional scientific analyses have two characteristics, they are made on a one-scene-at-the-time basis and they depend on human skills to manage and validate the classification of lots of files. Both characteristics not only discourage large-scale studies but also and make harder to validate and reproduce their results. Besides, the one-scene-at-the-time approach catches the spatial but ignores the temporal relations on the surface of the Earth. By contrast, time-series analysis allows to identify detailed characteristics of land cover change [7, 8, 9].

We believe the time series of a vegetation indexes of a tropical forests are regular and stable signals. In most situations, they show insignificant deviations. Therefore, the significant deviations in these stable signals are likely to be associated to deforestation events.

Here, we argue that the Kalman filter is an analysis tool which can discover deforestation events in large sets of EO time series of vegetation indexes. The Kalman filter has a well-established behaviour of self-correction, that is, it rapidly approximates the true observation value after a few observations. Deforestation or fire cause strong disturbances in the vegetation indexes that are identifiable as increasing differences between the observed and expected (filtered) values. The Kalman filter can process petabytes of EO data because of its reliability and low computational cost. For example, it was used for the guidance and navigation systems on board of the Apollo mission to the moon using just a

Thanks to the São Paulo Research Foundation (FAPESP) e-science program (grant 2014-08398-6). Gilberto Câmara is also supported by CNPq (grant 312151-2014-4).

computer with 36-bit floating-point arithmetic [10].

2. THE KALMAN FILTER

The Kalman filter is an adaptive filter which formulation addressed three problems: the prediction of random time series, the separation of the time series from noise, and the detection of well-know forms (pulses, sinusoids) in the presence of noise. It is implemented as an iterative process on which the outputs of one iteration are the inputs for the next one. In this way, the filter successively improves the estimations of the state of a system. Its inputs are the uncertainties in the observation and in the estimation, and the observations themselves; the outputs are the estimated true values of the system parameters. The Kalman filter process starts with the values of uncertainties in the observation and the estimation, and the observation itself. Then, a measurement of confidence — the Kalman gain — is calculated and it is used in the next step along an observation for estimating the true value of the observations (the observation value without noise). The Kalman filter is known for its fast approximation of the true values of the system parameters [11].

The extended Kalman filter (a non-linear Kalman filter) was used to detect new human settlements in the South African savanna. They used time series of Fourier-filtered MODIS data of 500-meter 8-day spatiotemporal resolution. The time series were modeled using cosine functions depending on the mean, amplitude, and phase. These three parameters were approximated — for each pixel — using the extended Kalman filter. A deviation of a pixel from its 8 neighbors is considered a land cover change [12]. Although its flexibility, this method is not intended for early-detection but for classification. Besides, its computational costs is increased due to its non-linear nature and its requirement for pre-processing (Fourier filtering). For that reason, we decided to use the Kalman filter instead of the extended version of it.

3. EXPERIMENT

For our experiment, we prepared a data set of vegetation indexes of 17 years of the Moderate Resolution Imaging Spectroradiometer (MODIS MOD13Q1) of the Amazon forest. MODIS data have a spatial resolution of 230 m and a temporal resolution of 16 days and their vegetation indexes are calculated from cloud-free, atmospherically corrected, and nadir-adjusted data [13]. Our data set had 253.4 million of time series, and each time series had more than 350 observations. The size of the compressed data set was 273 GB.

To hold the MODIS data set, we used a distributed cluster of the array database SciDB, which is an open source and optimized for of big data management. SciDB follows a *shared nothing* architecture paradigm on which the arrays are broken into *chunks* that are distributed among different servers; each server controls its local data storage and memory. SciDB

arrays are multidimensional and uniform, as each array cell holds the same user-defined number of attributes [14].

To run our experiment, we prepared an R script for processing a single time series and we let SciDB to distribute and execute our script in parallel. This approach enabled us to test our scripts on desktop machines before scaling them up to our SciDB cluster. This eases the writing-testing-debugging cycle and it has been used on similar studies [15, 16, 17].

Our script used MODIS metadata to remove observations of low quality or reliability. Then it filtered the time series of a forest's vegetation indexes using the Kalman filter. Our implementation of the filter filled in the missing observations using its own estimations. As the script ran, it computed the cumulative sum of the differences between the filtered and non-filtered time series. The differences above the three standard deviation threshold were classified as deforestation events.

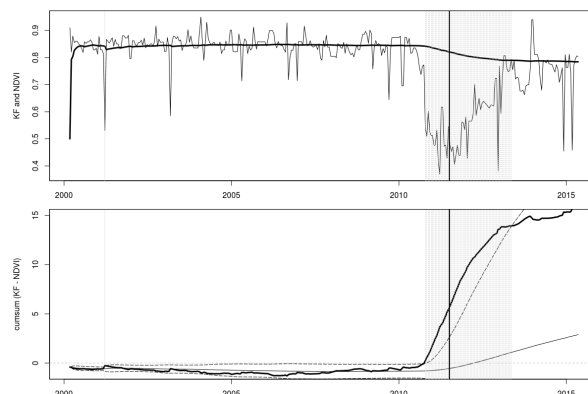


Fig. 1. Detection of deforestation using the Kalman Filter. The top panel shows an MODIS NDVI time series (thin line) and the Kalman filter (thick line). The bottom panel shows the cumulative sum of the KF - NDVI differences (thick line), the mean of the differences (thin line), and a three standard deviation threshold (thin dashed line). The deforestation observations are those where cumulative sum is outside the three standard deviation threshold (vertical dotted lines). The solid vertical line is the deforestation date reported by DETER.

To better illustrate our method, we use a time series of the Normalized Difference Vegetation Index (NDVI) in the town of *Pontes e Lacerda*, *Mato Grosso* state, Brazil at $14^{\circ}55'8.76''S$ and $59^{\circ}7'4.11''W$. The DETER monitor program reports this point as deforested in Tuesday 5th July, 2011. We filtered out the observations reported as low quality (e.g. no data, snow, ice, clouds, etc.); we also replaced those values by the latest valid observations. Then, we computed the mean and three standard deviations of the KF - differences. Any observation laying beyond the three standard deviation threshold is classified as deforestation (see Figure 1). In this particular instance, our method classifies observations as deforestation before the DETER monitor program and it continues to do it so as long as the cumulative

Table 1. Confusion matrix for each class using the the Kalman filter method. PA stands for producers accuracy and UA for user's accuracy.

	Deforestation	Forest	UA
Deforestation	2280	31595	0.06%
Forest	300	7272	0.96%
PA	0.88%	0.18%	

sum of the differences between the NDVI and the filtered time series falls outside the control threshold. For this reason, the detected deforestation period extends to several consecutive observations in contrast to the single date reported by DETER.

4. PRELIMINARY RESULTS

We applied our method to the town of *Manicoré*, Amazonas state, Brazil. The test area is bounded by the box defined by the WGS84 coordinates $8^{\circ}6'36.59''S$; $61^{\circ}48'21.96''W$ and $7^{\circ}37'40.87''S$; $61^{\circ}10'11.78''W$. We compared our results to DETER reports of deforestation (from 2014 to 2016) and the PRODES data reported as forest in 2016. It is worth mentioning that DETER only issues warnings about deforestation. For this reason, we used forest data from PRODES to check our method for misclassification. Additionally, the main concern of PRODES is accuracy while DETER is focused on accuracy and also speed. For example, it is possible to find in DETER duplicated deforestation warnings for the same place at different times as it is more prone to error than PRODES.

The results of our method are limited but they are still preliminary (see Table 1). As a reference, we also ran BFAST monitor on the same data, and the results are comparable (see Table 2). BFAST monitor is a change detection method based on the decomposition of time series of vegetation indexes into the trend, seasonal, and noise components [9].

These results suggest further pre-processing of MODIS data before applying our method as they are probably caused by the noise in the MODIS time series of vegetation indexes. This noise is due to clouds and sensor properties such as view angle and geometry [18].

Regarding execution time, our Kalman filter method and BFAST monitor took each approximately 2 milliseconds to analyze a single time series. However, our implementation of the Kalman filter is not optimized.

5. FINAL REMARKS AND FUTURE WORK

We propose an early warning method based on adaptive filtering — e.g. the Kalman filter — which has the potential to improve forest monitoring programs by increasing their capacity to process satellite images. However, we need to continue researching how to improve our results.

Table 2. Confusion matrix for each class using BFAST monitor. PA stands for producers accuracy and UA for user's accuracy.

	Deforestation	Forest	UA
Deforestation	2515	26803	0.08%
Forest	65	12064	0.99%
PA	0.97%	0.31%	

We are evaluating alternatives for future analysis, such as additional preprocessing of the MODIS time series or move to a more robust algorithm for the analysis such as the extended Kalman filter.

AN early detection can diminish deforestation by issuing timely warnings to authorities. Besides, the more data processed, the better the changes of detecting deforesters who hide from forest monitoring programs. This would have a positive impact in the emission of greenhouse effect gases due to land cover change.

6. REFERENCES

- [1] Robin L Chazdon, Pedro H S Brancalion, Lars Laestadius, Aoife Bennett-Curry, Kathleen Buckingham, Chetan Kumar, Julian Moll-Rocek, Ima Célia Guimarães Vieira, and Sarah Jane Wilson, "When is a forest a forest? Forest concepts and definitions in the era of forest and landscape restoration," *Ambio*, vol. 45, no. 5, pp. 538–550, sep 2016.
- [2] Jonathan a Foley, Ruth Defries, Gregory P Asner, Carol Barford, Gordon Bonan, Stephen R Carpenter, F Stuart Chapin, Michael T Coe, Gretchen C Daily, Holly K Gibbs, Joseph H Helkowski, Tracey Holloway, Erica a Howard, Christopher J Kucharik, Chad Monfreda, Jonathan a Patz, I Colin Prentice, Navin Ramankutty, and Peter K Snyder, "Global consequences of land use.," *Science*, vol. 309, no. 5734, pp. 570–4, 2005.
- [3] Matthew C Hansen, Stephen V Stehman, and Peter V Potapov, "Quantification of global gross forest cover loss," *Proceedings of the National Academy of Sciences*, vol. 107, no. 19, pp. 8650–8655, may 2010.
- [4] Gregory P. Asner, David E. Knapp, Eben N. Broadbent, Paulo J. C. Oliveira, Michael Keller, and Jose N. Silva, "Selective Logging in the Brazilian Amazon," *Science (80-.)*, vol. 310, no. 5747, pp. 480–482, 2005.
- [5] Peter Richards, Eugenio Arima, Leah VanWey, Avery Cohn, and Nishan Bhattarai, "Are brazil's deforesters avoiding detection?," *Conservation Letters*, pp. n/a–n/a, 2016.
- [6] Yosio Edemir Shimabakuro, João Roberto dos Santos, Antonio Roberto Formaggio, Valdete Duarte, and

- Bernardo Friedrich Theodor Rudorff, “The brazilian amazon monitoring program: Prodes and deter projects,” in *Global forest monitoring from earth observation*, Frédéric Achard and Matthew C Hansen, Eds., chapter 9, pp. 167–183. CRC Press, 2012.
- [7] Monya Baker, “Is there a reproducibility crisis?,” *Nature*, vol. 533, no. 7604, pp. 452–454, may 2016.
- [8] T. Blaschke, “Object based image analysis for remote sensing,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 1, pp. 2 – 16, 2010.
- [9] Jan Verbesselt, Achim Zeileis, and Martin Herold, “Near real-time disturbance detection using satellite image time series,” *Remote Sensing of Environment*, vol. 123, pp. 98–108, aug 2012.
- [10] Mohinder S. Grewal and Angus P. Andrews, “Applications of Kalman filtering in aerospace 1960 to the present,” *IEEE Control Systems Magazine*, vol. 30, no. 3, pp. 69–78, 2010.
- [11] R E Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35, 1960.
- [12] W Kleynhans, J C Olivier, K J Wessels, ..., B P Salmon, F van den Bergh, and K Steenkamp, “Detecting Land Cover Change Using an Extended Kalman Filter on MODIS NDVI Time-Series Data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, pp. 507–511, 2011.
- [13] R Solano, K Didan, A Jacobson, and A Huete, *MODIS vegetation indices (MOD13) C5 user’s guide*, Terrestrial Biophysics and Remote Sensing Lab of the University of Arizona, 1 edition, may 2010.
- [14] Michael Stonebraker, Paul Brown, Alex Poliakov, and Suchi Raman, “The architecture of SciDB,” in *23rd International Conference on Scientific and Statistical Database Management (SSDBM 2011)*, Judith Bayard Cushing, James French, and Shawn Bowers, Eds. 2011, vol. 6809 of *Lecture Notes in Computer Science*, pp. 1–16, Springer.
- [15] Gilberto Camara, Luiz Assis, Gilberto Queiroz, Karine Ferreira, Eduardo Llapa, Lubia Vinhas, Alber Sanchez, Victor Maus, and Ricardo Cartaxo, “Big Earth Observation Data Analytics: Matching Requirements to System Architectures,” in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. Association for Computing Machinery, 2016.
- [16] Meng Lu, Edzer Pebesma, Alber Sanchez, and Jan Verbesselt, “Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from {MODIS} time series,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 227–236, 2016.
- [17] Victor Maus, Gilberto Camara, Ricardo Cartaxo, Alber Sanchez, Fernando M Ramos, and Gilberto R. de Queiroz, “A Time-Weighted Dynamic Time Warping Method for Land-Use and Land-Cover Mapping,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3729–3739, aug 2016.
- [18] Jennifer N Hird and Gregory J McDermid, “Noise reduction of NDVI time series: An empirical comparison of selected techniques,” *Remote Sensing of Environment*, vol. 113, no. 1, pp. 248–258, 2009.

EO BIG DATA ANALYTICS FOR THE DISCOVERY OF NEW TRENDS OF MARINE SPECIES HABITATS IN A CHANGING GLOBAL CLIMATE

Sabeur, Z. A.¹; Correndo, G.¹; Veres, G.¹; Arbab-Zavar, B.¹; Neumann, G.¹; Ivall, T.¹; Castel, F.²; Zigna, J-M.³; Lorenzo, J.⁴

¹University of Southampton IT Innovation Centre, UK.

²Atos France, Toulouse, France.

³Collecte Localisation Satellite, Toulouse, France.

⁴Atos Spain, Madrid, Spain.

ABSTRACT

Climate change has been observed using multiple methods of Earth Observation (EO) including in situ, airborne and space-borne sensing methods. These use multimodal observation platforms, with various geospatial coverages, spatio-temporal resolutions and accuracies. The resulting EO Big Data from heterogeneous sources constitute valuable sources for scientists to investigate on the manifested responses of natural species behaviour to climate change. In the EO4wildlife¹ research project, we have access to Copernicus and Argos EO Big Data for conducting studies on the changes of habitats for a variety of marine species. The challenge is to discover causality of Metocean environmental observations and their relationship with the changing habitats of species. Nevertheless, there is a need to deploy Big Data technologies for connecting, ingesting, processing of EO data, as well as implementing specialised open data analytics services in this study. The particular services shall be made accessible to the scientific community for setting up modelling scenarios concerning the potential discovery of new trends of marine species habitats due to climate change. Three marine species are being studied in the EO4wildlife project. They include the Bluefin Tuna in the Atlantic-Mediterranean migratory regions, the black-footed albatross seabirds across the sub-tropical Atlantic Ocean and Loggerhead sea turtles along the North West coast of the African continent and Cape Verde. Large data representing geospatial migratory tracks and settlements of these respective marine species have been acquired in the project over period of times together with Metocean EO data from Copernicus and Argos satellites. These are currently analysed and modelled with a set of features obtained by searching in a large space of possible measured and derived Metocean parameters. A two-step search was used involving significance measurement and an iterative breadth first search based wrapper type feature selection algorithm. Furthermore, the analysis is useful for improving the performance of our habitat prediction models across the three marine species in the study. The discovery of new habitats geospatial and temporal trends which may be associated to the changing

climate under these analyses will be achieved through the deployment of web-enabled data mining and analytics open services. A dedicated Big Data platform supported by generic data management services in the cloud is therefore deployed for assuring the scalability of the data processing and analytics services.

Index Terms— Big Data, Earth Observation, Copernicus satellite, Climate change, habitat modelling

1. INTRODUCTION

EO4wildlife brings large number of multidisciplinary scientists such as marine biologists, ecologists and ornithologists around the world to collaborate closely together while using European Sentinel Copernicus Earth Observations more efficiently [1]. In order to reach such important capability, an open service oriented platform with an interoperable toolbox, that is compliant with OGC standards and supported by scalable cloud infrastructure is being implemented. The EO4wildlife platform offers dedicated open services that enable scientists to connect to marine species tracks databases and Big EO data in order to run habitat modelling simulations under a scalable processing environment. In particular, the platform enables the full integration of Copernicus sentinel data, ARGOS archive databases and animal track databases which can be effectively mined and fused for advanced big data analytics concerning the discovery of new trends of animal behaviour in the marine environments.

2. OPEN SERVICE ARCHITECTURE FOR BIG DATA MANAGEMENT

The EO4wildlife platform is composed of various functional components: 1- An internal data catalogue for aggregating geo-referenced products from external heterogeneous sources; 2- An ingestion module that allows the retrieval of data for exploitation by the platform services and; 4- A service Manager with which developers and/or data scientists manage the life cycle and execution of deployed services. Finally, the platform has built-in visualization features for the

¹ <http://eo4wildlife.eu/>

resulting geographic data from the processing services. The service management mechanism in the Big Data infrastructure is built on the containerization concept (i.e. Docker) which allows to encapsulate each service into an independent component that can be easily deployed on the cloud. An orchestration technology (i.e. Kubernetes) is used to manage container life cycle so that the underlying infrastructure becomes totally transparent [2].

3. BIG EO DATA ANALYTICS

In order to provide proofs of concept of the EO4wildlife platform and its dedicated Big EO data analytics services, a number of scenarios on habitat modelling for marine species behaviour are being developed. These required a pre-processing and analysis of the acquired big data for the discovery of strengths and relationships between data features prior to achieving efficiently performing models.

3.1. Big Data Features Selection

Features selection is the process of selecting the most dominant and connected variables or features for modelling environmental processes. Although initially there is only a small set of features (e.g. 8 features in the case of the pelagic fish use case) derived features such as gradients over time, averages over time and gradient over horizontal and vertical space are important to consider as they are related to the physical dispersion of nutrients and other hydrodynamic transport processes that take place within the marine environment [3]. In this case a genetic algorithm is used to search in the space of potential feature subsets. For each subset of selected features, ecological envelopes based on percentiles that the algorithm chooses and combines it into trees using “AND” and “OR” logical assertions are discovered. This process is performed stochastically and repeatedly so that a good number of possible subsets (therefore models) can be explored and trialled on the training set. Prior to this step a systematic search to find the best granularity for each derived feature (e.g. establishing whether temporal gradient for a given feature should be on a 10 or 30 day scale) is also conducted. This big data features selection process aims at optimizing the feature set to be used for best niche modelling the relationship between EO data and processes with trends on observed animal presence in space and time.

3.2. Habitat Modelling

Habitat niche modelling is a method for discovering and modelling the link between where the animal has been found (presence) and the environmental conditions at those points. These methods give an indication of the conditions which are favourable for the animal. Similarly, where the animals have not been found (absences) give an indication on the conditions that are not suitable for the animals. Given a model of climatic changes that forecasts metocean environmental

conditions, a habitat model for given species can be used to predict how the boundaries of its habitat do change due to such environmental conditions. One of the most concerning results of climate change is the vulnerability of habitats of certain species. Other problems may include rapid shifts in the spatial positioning of these habitats which can have severe consequences for less mobile species. In order to visualise the animal tracks as they evolve in time, and compare the distribution of metocean observations where the animals have been detected, a working demonstrator is being developed (see Figure 1). The demonstrator allows users to integrate and explore different types of data under a single user interface.



Figure 1 Spatial density distribution of sea turtles versus sea surface temperature environmental Observations

3.2.1. Habitat Modelling for Atlantic Bluefin Tuna (ABFT)

An Ecological Niche Modelling (ENM) framework which uses using observed animal presence data (animal tracks) has been developed for predicting probabilities of *Potential Habitats*. Specifically, monthly ENMs on *Potential Habitat* predictions of ABFT in the Mediterranean Sea were developed. The most relevant Earth Observation (EO) variables which influence habitat preferences were also identified [4]. These include Bathymetry, Sea Surface Temperature (SST), Chlorophyll (CHL), CO₂ Net Primary Production (NPP), Sea Level Anomalies (SLA) and Eddy Kinetic Energy (EKE). Environmental Envelopes (EE) were calculated during the model training stages for each variable through using pre-defined bounds. During the testing stage, geospatial areas of interest in the Mediterranean Sea were analysed with [0.1 x 0.1] degrees grid resolution. Each grid cell was set up to unity (*Potential Habitat* = 1), if for example, the sampled EO variables at the grid cell satisfies some specific environmental conditions, such as:

$$CHL_{min} \leq CHL_{(i)} \leq CHL_{max}, SST_{min} \leq SST_{(i)} \leq SST_{max}$$

The model for predicting *Potential Habitat*(0/1) is simply defined as follows:

$$Bathy_{range} (0/1) * SST_{range} (0/1) * CHL_{range} (0/1) * NPP_{range} (0/1) * EKE_{range} (0/1)$$

As a result, 99% percentiles for EE bounds were obtained. (See Table 1). *Proportion of Sea* notes the fraction of the spatial region that was classified as Habitat. *Number found* is the number of observed relocations that are considered as *Potential Habitat*. *Out of* is the number of all observed relocations. % in the last column is the percentage of correct predicted relocations in potential habitat.

Description	Proportion of Sea	Number found	Out of	%
ABFT habitat	0.679	80	85	94.12

Table 1. Potential Habitat Modelling for ABFT

3.2.2. Habitat Modelling for Black-Browed Albatross (BBA)

Though for the BBA species, only presence data are available, it is common practice to generate animal pseudo-absences techniques [5]. The generated pseudo-absences should be well separated from presences both in spatial and environmental (or ecological) space. The pseudo-absences are selected using a two-step approach. First, Correlated random Walk is used to generate 10 pseudo-absences for each presence relocation, where a constraint function is used to implement a spatial separation of presences from pseudo-absences. Second, EE and ENM is used to select the number of pseudo-absences which are well separated in environmental space. Though [6] performed a number of experiments and gave some recommendations on a number of pseudo-absences for different habitat modelling techniques, the experiments showed that equal number of presences and pseudo-absences lead to more robust performance for our Big data. Therefore we selected as many pseudo-absences as presences in the second step of pseudo-absence selection. This led us to a two-class problem for each geographic grid cells. Basically classified as either as *Potential Habitat* (=1) or *no Potential Habitat* (=0). Two regression techniques were used to predict *Potential Habitat* for the BBA. These include: A Generalised Additive Model (GAM) and Boosted Regression Trees (BRT). The EO data which influence *Potential Habitat* selections were in this case: *Bathymetry, SST, SLA and EKE*. The *Potential Habitat* modelling was done for each animal breeding stage (or monthly for non-breeding stage). The comparison of GAM and BRT for incubation stage both on training and testing set are given in Table 2, where Correct Classification Rates (CCRs) are shown for each class. The threshold for selecting habitat/no habitat was set to 0.5. Table 2 also shows that BRT produces better results both on training and testing modes.

Classifiers	Training		Testing
	Habitat	No Habitat	Habitat
GAM	77.1%	76.3%	68.45%
BRT	93.39%	99.51%	91.65%

Table 2. Correct Classification Rates(CCRs) for BBA

3.2.3. Habitat Modelling for Loggerhead sea turtles

Twenty one tracks of data on adult loggerhead sea turtles capturing their post-nesting movements during the years of 2004-2009 were also used for habitat modelling in this work. Two different foraging behaviours were observed with this animal population. These have been manually identified, and each animal was labelled as either an oceanic or a neritic forager. The overall modelling, pre-processing and pseudo-absence selection methods in this case were based on the works by Pikesley et al. [7], [8]. Three classification methods have been added and compared to the regression methods which were investigated in these works. Different spatial extents and numbers have also been examined for pseudo-absences. Data pre-processing stages include discarding relocations with unlikely speeds and turning angles. Best non-interpolated daily locations were then extracted for each of the tracks. Pseudo-absences were then generated within the convex hull of the presences via a random spatial-temporal sampling technique. Similar number of pseudo-absences as available presences were also generated (prevalence \geq 1). The post-nesting habitat for oceanic adult loggerhead sea turtles was modelled using different classification and regression models. These experiments on EO data were performed eight times (8 replications) using different random sets of pseudo-absences [9]. In each replication, the data is split with a 75%/25% ratio for training and validation purposes. This random data splitting to training and validation sets is independently repeated four times in each replication. Table 3 shows the modelling evaluation results using TSS (*True Skill Statistic*), which is the most widely used stat alongside kappa for evaluating the accuracy of the species distribution models [10], and AUC (*Area Under Curve*) as the only non-threshold based evaluation method. The reported results are the mean of all the performances in all the runs and replications. It can be seen that overall classification methods provide better models while they can be further improved by building ensemble models with the four runs in each replication.

	Regression models				Classification models		
	GLM	MARS	MAXENT	GAM	RF	BRT	CTA
TSS	0.347	0.496	0.414	0.434	0.606	0.536	0.531
AUC	0.722	0.813	0.767	0.770	0.875	0.843	0.814
TSS(EM)	0.314	0.480	0.424	0.419	0.922	0.545	0.667
AUC(EM)	0.712	0.818	0.785	0.766	0.995	0.856	0.893

Table 3. Habitat modelling results for Loggerhead sea turtles (GLM (Generalized Linear Model), MARS (Multiple Adaptive Regression Splines), MaxEnt (Maximum Entropy), GAM (Generalized Additive Model), RF (Random Forest), BRT (Boosted Regression Trees), CTA (Classification Tree Analysis). (TSS (EM) and AUC (EM) are ensemble models based on the four runs in one replication.)

4. SUMMARY

The above research work used Copernicus and Argos Big data resources while adopting Big data infrastructure for wrapping a new generation of big data analytics services for predicting marine species habitats. The main focus of the EO4wildlife project was to establish performing mining and data analytics methods which automatically extract new knowledge from the newly available Copernicus Big EO data combined with those from Argos. The extracted knowledge, specifically concerns the confirmation of existing causalities between new emerging ecological conditions, due to climate change, and the response of selected vulnerable animal species at various oceanic regions. The next activity in the project will be on validating the elasticity and scalability of the big data analytics services which are being implemented on the EO4wildlife platform in collaboration with our project partners.

5. ACKNOWLEDGEMENT

This work is partly funded by the European Commission under H2020 Grant Agreement number: 687275. Also, the Authors would specifically like to thank the European Commission for providing access to COPERNICUS data, under the EO4life project, without which this study could not have been conducted.

6. REFERENCES

[1] Z. Sabeur, G. Correndo, G. Veres, B. Arbab-Zavar, J. Lorenzo, T. Habib, A. Haugommard, F. Martin, J-M. Zigna, and G. Weller. (2017) *EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife*. Proceedings of the 12th International Symposium of Environmental Information

Systems. Zadar, Croatia, May 10th – 12th 2017. Springer International Publishing.

[2] G. Correndo, J-M Zigna, A. Haugommard. (2016). D3.1: Knowledge Base Service architecture Specification v1. Deliverable of EO4wildlife project. (see <http://www.eo4wildlife.eu/deliverables> - [WP3 Advanced Analytics and Knowledge Base](#))

[3] Druon, Jean-Noël, et al.(2011). *Potential feeding and spawning habitats of Atlantic bluefin tuna in the Mediterranean Sea*. Marine Ecology Progress Series 439: 223-240.

[4] J.-N. Druon and et al. (2016) Habitat suitability of the Atlantic bluefin tuna by class size: An Ecological niche approach. Progress in Oceanography, vol. 142, pp. 30-46.

[5] S.D. Senay, S.P. Worner, T. Ikeda (2013). Novel three-step pseudo-absence selection technique for improved species distribution modelling. PLoS ONE, p. e71218

[6] Barbet-Massin M, Jiguet F, Albert CH, Thuiller W (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3: 327–338.

[7] S. K. Pikesley, S. M. Maxwell, K. Pendoley, D. P. Costa, M. S. Coyne, A. Formia and S. Ngouesso. (2013) "On the front line: integrated habitat mapping for olive ridley sea turtles in the southeast Atlantic. "Diversity and Distribution". Vol.19, no 12, pp.1518-1530.

[8] S. K. Pikesley, A. C. Broderick, D. Cejudo, M. S. Coyne, M. H. Godfrey, B. J. Godley and M. J. Witt. (2015). "Modelling the niche for a marine vertebrate: a case study incorporating behavioural plasticity, proximate threats and climate change," *Ecography*, vol. 38, no. 8, pp. 803-812.

[9] M. Barbet-Massin, F. Jiguet, C. H. Albert and W. Thuiller. (2012). "Selecting pseudo-absences for species distribution models: how, where and how many?," *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 327-338.

[10] O. Allouche, A. Tsoar and R. Kadmon. (2006). "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)," *Journal of applied ecology*, vol. 43, no. 6, pp. 1223-1232.

SOCIOECOLOGICAL CARBON PRODUCTION IN MANAGED AGRICULTURAL-FOREST LANDSCAPES

Jiquan Chen¹, Kyla Dahlin¹, Ranjeet John¹, Gabriela Shirkey¹, Susie R. Wu¹, Phil Robertson¹, Steve Hamilton¹, Lauren Cooper¹, Dave Lusch¹, Arnon Karnieli², Raffaele Laforteza^{1,3}, Giovanni S. Labini⁴, Angelo Amodio⁴

¹Michigan State University - USA; ²Ben-Gurion University of the Negev - Israel; ³University of Bari - Italy, ⁴Planetek - Italy

ABSTRACT

Land use, land cover changes, and ecosystem-specific management practices are recognized for their roles in mediating the climatic effects on ecosystem structure and function, (e.g., C cycle). A **major challenge** is to understand and forecast ecosystem C fluxes, we cannot rely solely on conventional biophysical regulations from the local ecosystem to the global scale. A **second challenge** is to quantify the magnitude of the C fluxes from managed ecosystems/landscapes over the lifetime of the C cycle, and to deduct the various energy inputs during management. The **objective of this project**, started in Spring 2017 and funded by the **Carbon Cycle & Ecosystems (CC&E)** of the **NASA Earth Science Program**, is to quantify the landscape-scale C footprint of both managed agricultural-forest landscapes and people.

The three **fundamental questions** are: 1) what are the quantitative contributions of land cover change, specific management practices, and climate changes to the social and physical C fluxes of managed ecosystems and landscapes; 2) what are the spatial and temporal changes of their contributions in managed agricultural-forest landscapes; and 3) how will future land use changes impact C sequestration in an upper, mid-latitude managed ecosystem?

Hypothesis:

Social C flux is more responsible than physical C flux for the dynamics, and especially the uncertainty, of the cumulative CO₂^{eq} production of these intensively-managed landscapes. Their proportions vary among the landscapes and over history because of the great variations in land conversions, land use practices, climatic changes and extremes in the watershed.

Index Terms—carbon cycling, land use, flux towers, climate change

1. INTRODUCTION

Land use, land cover changes, and ecosystem-specific management practices are increasingly recognized for their roles in mediating the climatic effects on ecosystem structure and function. The scientific community on carbon (C) cycling, for example, has gained the much needed knowledge, technology, and tools to model and forecast the changes of C fluxes through a combination of large in situ measurements (e.g., FLUXNET, Papale et al 2015), remote sensing (RS) technology (e.g., Xiao et al 2012), and ecosystem models (e.g., Hurtt et al 2011, Schaefer et al 2012). Depending on the region, some scholars have demonstrated that human activities influence C fluxes and storage far more than climatic changes (IPCC 2014). The UNFCCC Conference of the Parties, which formed to address climate change, signed the historic “Paris Agreement” proposed by the U.S. in April 2016 along with nearly 200 other countries. The agreement prominently features land and forest sectors as important

carbon pools and opportunities for sequestration (The United Nations, Framework Convention on Climate Change 2015). Over the last decade, other policies and programs attempting climate mitigation, such as Reducing Emissions from Deforestation and Degradation (REDD), have been launched and scrutinized by academia. This study aligns well with the SDG, Space Economy and CCI+.

2. CHALLENGES

A **major challenge** is that our understanding and forecasting of ecosystem C fluxes cannot rely solely on conventional biophysical regulations at any scale, from the local ecosystem to the globe. A **second challenge** is to quantify the magnitude of the C fluxes from managed ecosystems and landscapes over the lifetime of the C cycle, and to deduct the various energy inputs during management from the amount of C sequestered by an ecosystem (West & Marland 2003). For example, conventional crop management often includes tillage, fertilization, irrigation, applications of pesticides and herbicides, harvesting, transportation to the market, land conversion, etc. All of these activities require a CO₂-equivalent (CO₂^{eq}) amount of energy (hereafter “**social C flux**”) to offset the actual amount of C sequestered by the ecosystems and landscapes. A complete life cycle assessment (**LCA**) is needed to account for the actual sequestration strength at different spatial and temporal scales. Our recent LCA study on converting marginal lands to biofuel systems in southwest Michigan indicated that, if the land was tilled, a C balance cannot be reached until 89–123 years pass (Gelfand et al 2011). When developing conceptual and predictive models to realistically quantify the C flux in time and space, one must also include human activity.

3. OBJECTIVES

Our **overall objective** is to quantify the landscape-scale C fluxes at annual scale of both managed agricultural-forest landscapes and people, using the Kalamazoo watershed in southwestern Michigan as our testbed. The underlying mechanisms from both human activities and biophysical changes (e.g., climate, phenology) on ecosystem C dynamics at different temporal and spatial scales are explored by modeling net ecosystem C production (hereafter “**physical C flux**”), estimating *social C flux*, exploring the complex relationships through Bayesian structural equation modeling, and performing a spatially-explicit LCA on the total C production within the contrasting landscapes and the entire watershed. The three **fundamental questions** are: (1) what are the quantitative contributions of land cover change, specific management practices, and climate changes (means and extremes) to the social and physical C fluxes of managed ecosystems and landscapes; (2) what are the

spatial and temporal changes of their contributions in managed agricultural-forest landscapes; and (3) how will future land use changes (including alternative management practices) impact C sequestration in an upper, mid-latitude managed ecosystem?

We take a bottom-up approach to quantify landscape C fluxes and a top-down scaling effort to characterize the contributions of climatic change, land use, and site-specific management practices at two spatial scales: landscapes in the Midwest region of the USA with contrasting structure and composition and the entire watershed (Fig. 1). Our **overarching hypothesis** is that *social C flux is more responsible than physical C flux for the dynamics, and especially the uncertainty, of the cumulative CO₂^{eq} production of these intensively-managed landscapes. However, their proportions vary significantly among the landscapes and over history because of the great variations in land conversions, land use practices, climatic changes and extremes in the watershed.*

Physical C flux of an ecosystem is jointly regulated by the biophysical conditions such as climate, soil, and vegetation. Supported by rich *in situ* measurements in diverse ecosystems, manipulative experiments, and ecosystem models from the past four decades, the magnitude and dynamics of the physical C flux can be predicted with high confidence. Application of RS technology in recent decades has also greatly advanced our predicative ability at broader spatial and temporal scales. Relatively underdeveloped, yet rapidly evolving, considerations are the roles of land use, management practice, and extreme climate that may significantly change or even tip the balance of C production. More importantly, at the landscape scale where different types and sizes of ecosystems exist, management strategies and actions play a leading role in determining the cumulative C production level (Chen et al 2004)

Social C flux refers to the amount of CO₂^{eq} that is transported among the ecosystems within the landscape (i.e., internal import and export), or from/to outside of the landscape (i.e., external import and export). Within the landscape, CO₂^{eq} transportation may include the transportation of organic materials, water, and nutrients that all need energy (i.e., CO₂^{eq} loss from the system). The external losses may involve much more CO₂^{eq}, such as harvesting, thinning of forest plantation, grain/feedstock or timber production, fuel burning, fertilization, irrigation, tillage (energy consumption), applications of pesticides and herbicide, etc. The amount of CO₂^{eq} involved in these activities needs to be deducted from the new landscape CO₂^{eq} balance. Depending on the management type and intensity, they can be significant.

When the physical C flux from land conversion, tillage, fertigation, harvesting energy consumption, and applications of pesticides, was deducted the CO₂^{eq} production of all three biofuel systems was negative for many years after initial establishment (Gelfand et al 2011, Zenone et al 2013). These social C fluxes are normally not considered or measured in C-cycle studies, but can be estimated through calculating the amount of energy or materials involved in each activity.

The CO₂^{eq} production of an ecosystem and the landscape (i.e., social C flux + physical C flux) also needs to be estimated through its life cycle because of the complex combination of practices and product uses over time and space. A detailed LCA is therefore needed to describe possible future situations that are relevant for specific management practices. Due to the scope and goal of this study, we only calculate the potential amount of CO₂^{eq} using the Product Life Cycle Accounting and Reporting Standard (Wu et al 2014). LCA is used to quantify the total CO₂^{eq} over different temporal scales and under alternative management/climate scenarios.

STUDY SITE

The Kalamazoo River watershed (5261 km²), which includes

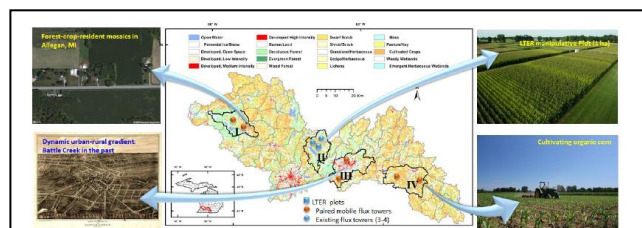


Fig. 2. Current land cover of the Kalamazoo Watershed (NLCD), which includes 127 subwatersheds (USGS). The entire watershed is examined for the changes of CO₂^{eq} during a 40-year period (1978–2018) using Landsat/Sentinel with the climate and human activities following our working framework (Fig. 1), while four contrasting landscapes are quantified with high-resolution RS data and historical records and survey statistics over an 80-year period (1938–2018).

portions of 11 counties (Allegan, Ottawa, Van Buren, Kent, Barry, Kalamazoo, Calhoun, Eaton, Jackson, and Hillsdale) in southwestern Michigan, is currently dominated by cultivated crops (32.9%), deciduous forest (20.0%), pasture-hay prairies (15.1%), lakes and wooded wetlands (14.7%), and urban areas (6.8%). After preliminary analysis of the landscape structure, we conduct our study in four contrasting landscapes within the watershed (Fig. 2). These landscapes are typical of those found in the US Midwest, which include: (I) heavily forested sub-watersheds along with seasonally flooded forest-wetland matrix at the mouth of the river near the Allegan Game Area; (II) the lake-forest-cropland mosaics around Gull Lake in Kalamazoo/Barry counties; (III) urban-rural gradients centered in Battle Creek; and (IV) cropland-dominated agricultural landscapes (Hamilton et al in review). These landscapes have an area of 20, 681; 19,504; 25,496; and 27,561 ha, consisting of 2, 3, 4, and 4 sub-watersheds (total=13) (Fig. 2).

4. RESEARCH TASKS

We follow three tasks to achieve our objectives. First, Landsat, Sentinel, MODIS, aerial photos, Worldview I-3, VENμS, AMSR-E and other spatial databases of biophysical and socioeconomic variables are compiled, processed and made accessible on our project webpage at the LEES Lab of Michigan State University to develop a comprehensive database of land use and land cover change for the watershed and landscapes over space and time. Then both a biophysical model (CLM) and a socioeconomic model are customized with intensive field data including vegetation, soil, climate, surveys, government statistics, as well as the RS land surface properties. We perform an uncertainty analysis by integrating Bayesian modeling with SEM. Through our collaborations with the LTER/GLBRC at KBS, the RS&GIS Center, and Planetek, the three research tasks provide an understanding of the changes and regulations of CO₂^{eq} production in the four contrasting landscapes (1938–2018) and the entire watershed (1978–2018). Ground measurements of C flux and potential drivers are crucial to model parameterization and validation. In addition to pooling the data from the seven current EC towers, we deploy a pair of mobile EC towers in contrasting landscapes to take periodic snapshot measurements, lasting 4-5 weeks each, in different land-cover types for independent model validation. In addition to the major research tasks, we provide outreach and engagement activities

through an interactive web interface, downloadable data and project results, and multi-media project updates, and to incorporate students into project activities.

We provide in the following details of Task 1 and a summary of Task 1 and 2.

4.1. Task 1: Dynamics of Physical C Fluxes

Objective: We aim to quantify the changes of the physical C flux on an annual scale, which are converted to CO₂^{eq}, by integrating: (1) remotely-sensed land cover type and other surface properties; (2) geospatial records of climate, vegetation, soil, and management practices for model parameterization; (3) direct measurements of net ecosystem exchange of CO₂^{eq} using EC flux towers for model validation; and (4) a customized ecosystem model (i.e. CLM).

Task 1.1. Land cover change (LCC) and land surface properties:

Land cover maps are created at two spatial resolutions: watershed (30 m) and landscape (1–4 m). Landsat/Sentinel are used to develop accurate land cover maps of the entire watershed for the 1978–2018 period. We acquire 30 m resolution Landsat scenes at 5-year intervals using the Global Landsat Survey (GLS) for years 1975, 1990, and 2010 and MSS/TM/OLI scenes for years 1980, 1985, 2000, and 2018—totaling a 40-year study period. We also characterize the present day landscape by using 10 m and 20 m ESA Sentinel-2 imagery to cross-check the classification and fill the data gaps. Existing classification products such as the 30 m NLCD might omit vegetation cover at fine scales, such as isolated forest stands and narrow strips of cropland cover. Therefore, hyperspatial data at 1–4 m resolutions from the National Geospatial Agency (NGA) commercial archive database are used to classify and archive the dominant agricultural, forest cover and urban green cover using Object based Classification. These data might include Worldview-3, Worldview-2, Worldview-1, and the QuickBird series in the present day as well as the legacy IKONOS. The NGA data enable a fine scale analysis of the percentage of canopy cover at stand level and crop type and productivity at the parcel level. To ensure a degree of consistency among all images, standard methods (Estoque & Murayama 2015, Poursanidis et al 2015) for land cover classification is consistently applied to all the scenes.

The long-term cover changes of the four contrasting landscapes is quantified with high resolution aerial RS data. There have been major proportional changes in forest cover relative to cropland cover in southwest Michigan. These important land cover/use changes at the landscape level can be only tracked back to the early Landsat/MSS era (1972). In order to classify and archive fine-scale changes in landscape structure, prior to the IKONOS era, we use aerial photos from 1938, 1958 and 1978, which are provided from the RS&GIS Aerial Imagery Archive. The more recent Digital Orthophoto Quadrangles (DOQs) from Michigan DNR data portals are used for 1992–1998. These fine scale aerial photos (e.g., RF 1: 14,000, 1:20,000 and 1:6,000) are used to identify cover type. While color infrared photography can be readily classified through standard image processing, we extract information from the black and white aerial photos through texture analysis and visual photo-interpretation.

In order to measure fine-scale changes in land cover and land surface properties on the four landscapes, we make use of the imagery at a high repetition (two days) from the Vegetation and Environmental New Micro Spacecraft (VENμS) at 5.3 m resolution and 27 km swath. The high observation frequency is essential for detecting the

dynamics of vegetation growth, the short duration of phenological stages, and the rapid temporal changes of water quality.

Other land surface properties: A suite of satellite-derived products including proxies of productivity (e.g., MODIS NDVI, EVI, LAI), water content indices (e.g., LSWI, NDSVI), land surface temperature (LST), soil moisture (AMSR-E, SMAP), and precipitation (TRMM, GPM) are used to develop a continuous database with high temporal resolution with wall-to-wall coverage at the regional and landscape hierarchies. Landsat/Sentinel (1978–2018) scenes and VENμS are used to quantify land surface properties (e.g., fraction of vegetation cover, land surface temperature, and above ground biomass) developed by scaling up from field sampling to the entire region. These remote-sensing products are used to explore the interactions among the various elements of the study as well as for model parameterization. We obtained the 8-digit and 12-digit hydrologic unit watershed boundaries prepared by the Michigan Department of Environmental Quality (MDEQ). Soil Survey Spatial and Tabular Data (SSURGO 2.2) for the Counties of Allegan, Barry, Calhoun, Eaton, Jackson, and Kalamazoo are obtained from the USDA NRCS portal, and a 10-m resolution DEM are obtained from the USGS.

Task 1.2: Climate and Soil Data: Daily mean, minimum, and maximum temperature, relative humidity, radiation, and precipitation from NOAA stations for the region are included in this database.

Task 1.2: Climate and Soil Data: Seven EC flux towers operates in the watershed since 2009 and are continuously maintained. A pair of mobile, open-path EC flux carts (2–3 m in height, available at the LEES Lab) are used and moved among different ecosystems to directly measure the net ecosystem exchanges (NEE) of carbon (CO₂, CH₄), water, energy, and microclimate.

Task 1.4. Modeling Physical C Fluxes: CLM is the land component of the Community Earth System Model (CESM) (Hurrell et al 2013). The CLM and CESM are state-of-the-art models that are developed and utilized by organizations and investigators around the globe. The CLM includes representations of the biogeophysical and biogeochemical processes that both influence and are influenced by the climate and ocean systems. The processes in CLM relevant to this project include surface energy balance, soil and snow temperatures, hydrology, stomatal resistance and conductance, an urban model, C and Nitrogen (N) cycling, phenology, decomposition, crops, and land cover change.

4.2. Task 2: Dynamics of Social C Fluxes

Objective: We aim to estimate the social C fluxes of major management practices for different land cover types by classifying historical land cover, identifying land ownership, and by surveying historical management practices of individual land-owners (parcel scale). Back-of-the-envelope calculations is applied to scale up the CO₂^{eq} fluxes to the landscapes and the watershed.

4.3. Task 3: The dynamics and the regulations of CO₂^{eq} in time and space

Objective: We aim to diagnose the mechanistic/empirical causal relationships based on biophysical models and SEM, and to quantify the ecosystem, landscape, and watershed C fluxes at multiple temporal scales and under alternative management/climate scenarios.

5. REFERENCES

- Papale, D., Black, T. A., Carvalhais, N., Cescatti, A., Chen, J., Jung, et al. (2015). Effect of spatial sampling from European flux towers for estimating carbon and water fluxes with artificial neural networks. *Journal of Geophysical Research: Biogeosciences*, 120(10), 1941-1957.
- Xiao, J., Chen, J., Davis, K. J., & Reichstein, M. (2012). Advances in upscaling of eddy covariance measurements of carbon and water fluxes. *Journal of Geophysical Research: Biogeosciences*, 117(G1), n/a-n/a. doi:10.1029/2011JG001889
- Hurtt, G. C., Chini, L. P., Frolking, S., Betts, R. A., Feddema, J., Fischer, G. et al. (2011). Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands. *Climatic Change*, 109(1), 117-161.
- Schaefer, K., Schwalm, C. R., Williams, C., Arain, M. A., Barr, A., Chen, et al. (2012). A model-data comparison of gross primary productivity: Results from the North American Carbon Program site synthesis. *Journal of Geophysical Research: Biogeosciences*, 117(G3), n/a-n/a. doi:10.1029/2012JG001960
- West, T. O., & Marland, G. (2003). Net carbon flux from agriculture: Carbon emissions, carbon sequestration, crop yield, and land-use change. *Biogeochemistry*, 63(1), 73-83.
- Gelfand, I., Zenone, T., Jasrotia, P., Chen, J., Hamilton, S. K., & Robertson, G. P. (2011). Carbon debt of Conservation Reserve Program (CRP) grasslands converted to bioenergy production. *Proceedings of the National Academy of Sciences*, 108(33), 13864-13869.
- Chen, J., Brosofske, D. K., Noormets, A., Crow, R. T., Bresee, K. M., Le Moine, et al. (2004). A working framework for quantifying carbon sequestration in disturbed land mosaics. *Environmental Management*, 33(1), S210-S221
- Zenone, T., Gelfand, I., Chen, J., Hamilton, S. K., & Robertson, G. P. (2013). From set-aside grassland to annual and perennial cellulosic biofuel crops: Effects of land use change on carbon balance. *Agricultural and Forest Meteorology*, 182–183, 1-12
- Wu, R., Yang, D., & Chen, J. (2014). Social life cycle assessment revisited. *Sustainability*, 6(7), 4200.
- Estoque, R. C., & Murayama, Y. (2015). Classification and change detection of built-up lands from Landsat-7 ETM+ and Landsat-8 OLI/TIRS imageries: A comparative assessment of various spectral indices. *Ecological Indicators*, 56, 205-217
- Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, et al. (2013). The community Earth system model: A framework for collaborative research. *Bulletin of the American Meteorological Society*, 94(9), 1339-1360.

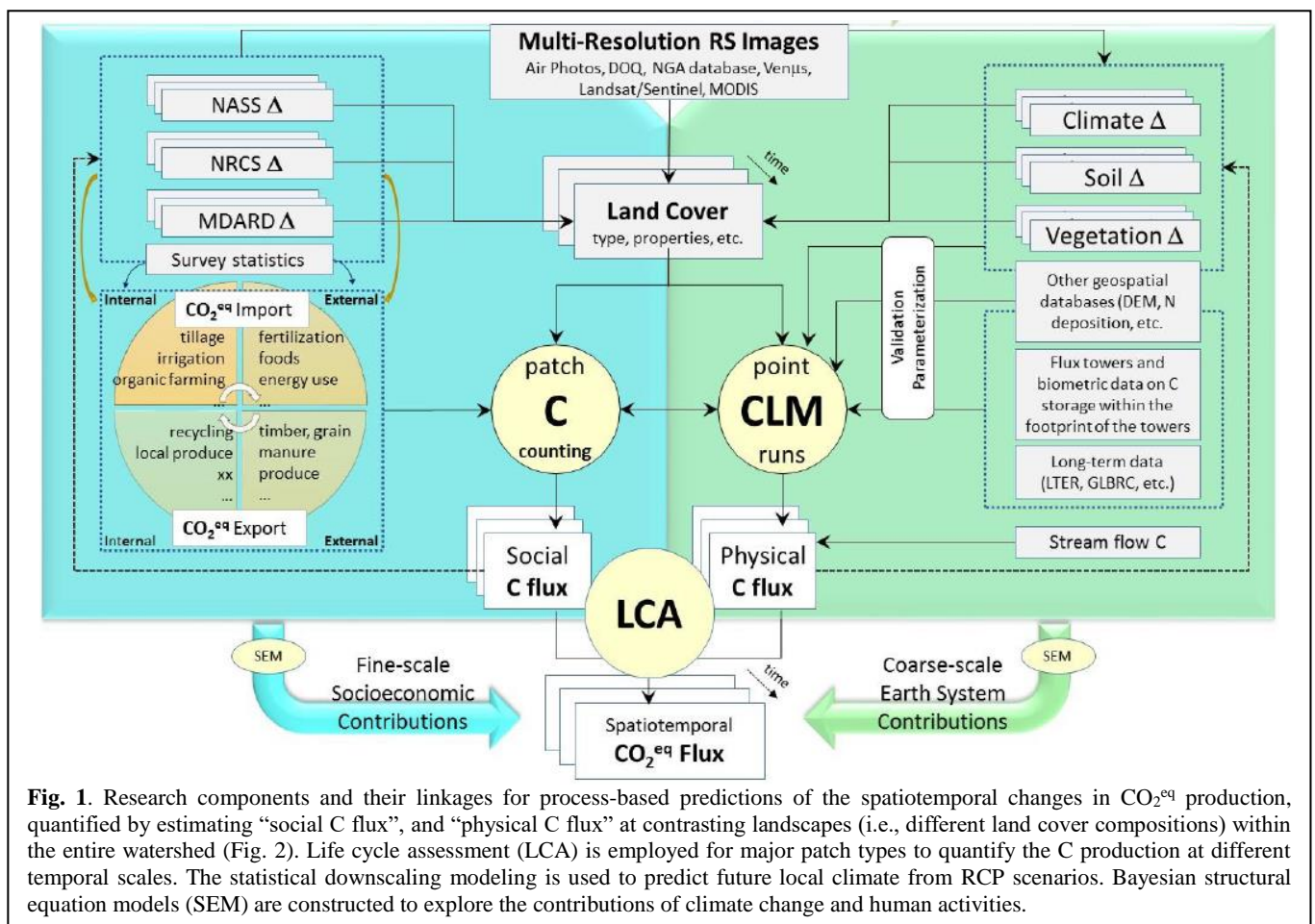


Fig. 1. Research components and their linkages for process-based predictions of the spatiotemporal changes in CO_2^{eq} production, quantified by estimating “social C flux”, and “physical C flux” at contrasting landscapes (i.e., different land cover compositions) within the entire watershed (Fig. 2). Life cycle assessment (LCA) is employed for major patch types to quantify the C production at different temporal scales. The statistical downscaling modeling is used to predict future local climate from RCP scenarios. Bayesian structural equation models (SEM) are constructed to explore the contributions of climate change and human activities.

THE COASTAL WATERS RESEARCH SYNERGY FRAMEWORK, FOR UNLOCKING OUR POTENTIAL FOR COASTAL INNOVATION GROWTH

Miguel Homem⁽¹⁾, Nuno Grosso⁽¹⁾, Nuno Catarino⁽¹⁾, Rory Scarrott⁽²⁾, Eirini Politi⁽²⁾, Abigail Cronin⁽²⁾

(1) Deimos Engenharia S.A., (2) University College Cork

ABSTRACT

Coastal and Ocean research communities are ill-equipped to familiarise themselves with, and take full advantage of, rapidly expanding Earth Observation (EO) data availability. A second consequence of this expansion in data availability is the difficulties encountered downloading and processing data. The H2020-funded Co-ReSyF project is a 3 year initiative to develop a cloud platform, which not only simplifies integration of EO data use into multi-disciplinary research activities, but also avoids requiring users to download large quantities of data, whilst educating and engaging non-EO communities in the benefits of integrating EO-derived data into their research. This platform aims to be user friendly and accessible to non-EO scientists as well as EO and coastal experts. The development of this platform is achieved iteratively in close consultation with a multi-level stakeholder community shaping not only the system usability, and tools range, but also the strategic focus of all the project's activities. This paper provides an overview of the projects developing system, thematic focus, community engagement and iterative stakeholder-led development activities.

Index Terms— Coastal, marine ecology, remote sensing, data fusion, collaborative, cloud, tools

1. INTRODUCTION

Until recently, scientists had to deal with the daunting task of mining large datasets for suitable data, and often downloading EO information from various different sources. Also, as the datasets increased in volume, the processing has become slower and more demanding of computing facilities. The Coastal Waters Research Synergy Framework (Co-ReSyF) project aims to tackle these issues, by developing a platform for combined data access, processing, visualisation and output in one place. The platform is based on cloud computing to maximise processing effort and task orchestration. It will support researchers in the field of monitoring the socioeconomic coastal activities (e.g. fisheries, harbour operations, ship traffic monitoring, oil spill detection) in a changing world. Similar to Co-ReSyF, other initiatives exist to tackle the problem of processing large EO datasets and the easy access to them. Initiatives like the ESA TEPs [3], EVER-EST [4], EO4WILDLIFE [5], EO4ATLANTIC [6] and

NEXTGEOSS [7]. The difference between Co-ReSyF and these other initiatives is its focus on animating research and education activities for a specific thematic area, the coastal area.

Co-ReSyF is a 3-year project (2016-2018) funded by the European Union to support the development of research applications using Earth Observation (EO) data for Coastal Water Research. Co-ReSyF will create a cloud platform, which simplifies integration of EO data use into multi-disciplinary research activities that fits the needs of inexperienced scientists as well as EO and coastal experts. We will reach a wide community of coastal and oceanic researchers, who are offered the opportunity to experience, test and guide the development of the platform, whilst using it as a tool for their own research. The platform will include a set of 5 core Research Applications, developed under the project, and also a set of tools that the researchers can use to build their own applications in a user friendly manner. Each of these research applications consists of subcomponent modules, which users can apply to different research ventures. Additionally, other potential tools or applications can be added by the research community for sharing with other researchers that may find it useful. The set of core applications to be developed during the project lifetime are:

- Bathymetry Determination from SAR Images
- Determination of bathymetry, benthic classification and water quality from optical sensors
- Vessel and oil spill detection
- Time-series processing for hyper-temporal optical data analysis
- Ocean coastal altimetry

Additionally, a group of 8 Master/PhD students have been selected to attend a Summer School where they learn how to use the platform and will also contribute with their own tools and/or applications to be incorporated into the platform.

2. THE CONSORTIUM

The consortium consists of eight partners, based in five different countries. It comprises a varied spectrum of entities ranging from academic institutions and national laboratories, to private companies and SMEs. Within the consortium we gather the technical expertise not only to develop the Co-ReSyF system, but also to actively test it with specific research applications, and generate a vibrant

multi-disciplinary research community who are using the system to pursue coastal waters research. The team is managed by DEIMOS Engenharia S.A. (Portugal), and is composed of University College Cork (Ireland), Terradue Srl. (Italy), ACRI-HE (France), National Oceanography Centre (UK), ARGANS Ltd. (UK), LNEC (Portugal) and Instituto Hidrográfico (Portugal).



Fig. 1: Consortium

3. CO-RESYF CLOUD PLATFORM

The project aims to tackle the problems faced by the researchers when using EO data, by developing a platform for combined data access, processing, visualisation and output in one place. Those components will be complemented by a set of user support systems that helps guiding the researcher through the wide array of datasets, applications and processing chains, available to him in the platform to achieve his research goals. The platform is based on cloud computing to maximise processing effort and task orchestration. Co-ReSyF will address issues faced by inexperienced and new EO researchers, and also target EO experts and downstream users. The main focus is on enabling EO data access and processing for coastal and marine applications. In a collaborative research environment, the Co-ReSyF platform will revolutionise accessibility to Big EO Data for EO and non-EO scientists, and help create a new era of EO data processing and exploitation in the coastal and marine environment research area.

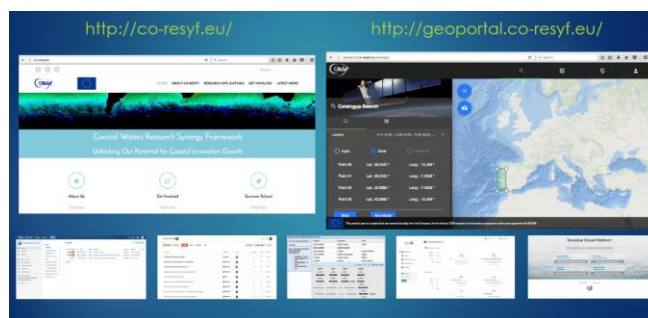


Fig. 2: The Co-ResyF Platform

3.1. The module concept

A module within the Co-ReSyF platform is a complete processing step, that can be used independently for data (pre-)processing (e.g. land masking, radiometric corrections, etc.), or combined with other modules to create processing chains required for the deployment of entire research applications. Modules can be combined to suit users' applications requirements. On one hand, inexperienced users can use the modules as they are, to build processing chains without the need to understand the underlying process/algorithm of the module. On the other hand, advanced users will be able to use the Co-ReSyF functionalities to develop their own modules, or configure copies of existing ones. The original module, managed on the platform as a template, will remain unchanged for future use. All datasets, modules and processing chains will be documented through their metadata so that they are traceable and replicable by other users. The metadata information will also be used

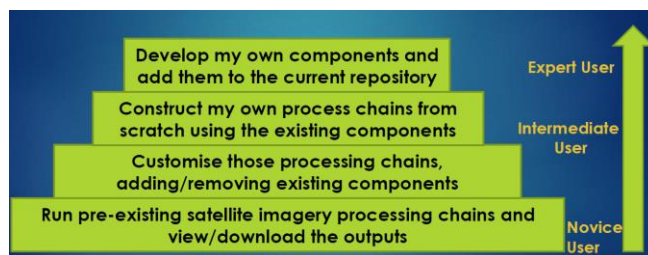


Fig. 3: Different User Scenarios

For dealing with the concept of modules the platform has the following definitions:

- **Tools:** Image/data processing units that the user can use to compose a workflow.
- **Workflow:** The representation of a processing chain with the ordered steps (i.e. tools) required to process the input data to obtain a specific output.
- **Workflow composer:** User interface where the user can create/edit workflows using the tools from the platform.
- **Workflow engine:** The backend component responsible for the execution of the processing

steps defined in a workflow in the cloud environment.

3.2. The support for non-expert users

The Co-ReSyF platform not only aims at supporting expert researchers in EO data to do their work, but also to help and educate non-experts in EO data which want to use it for their research. With that goal in mind the platform plans to provide a set of tools to assist the researchers in finding out what they need to do to derive the required data from the platform. These tools comprise of the following:

- **Knowledge Base:** Wiki like website where all technical and scientific information regarding the tools and applications integrated in the Co-ReSyF platform.
- **Expert Centre:** Automated rule based system integrated in the workflow composer that will provide guidance on each workflow component usage to the user when he is editing a workflow. It takes advantage of the metadata attached to each dataset, module and processing chain to define the usage rules
- **User Forum:** Online forum organized by categories where the platform users, developers and experts can discuss all issues related to the integrated functionalities
- **Whiteboard:** A platform component where the users can post ideas for research activities and other users can collaborate with them to define and implement the processing chains to derive the needed data for the research.
- **User Support Service Desk:** A user support system that will allow users to issue tickets to the development team reporting problems/suggestions on the platform.

3.3. Tools offered by the platform

The Co-ReSyF platform will offer a set of tools for handling EO data. Tools for processing functions like atmospheric corrections, cloud flagging, image inter-calibration, radiometric corrections, optimal interpolation and other more general functionalities like format conversions and tiling and aggregation of images. These tools are based on open source tools, e.g. SNAP and GDAL, and are mostly written in Python language. The image below shows the current status of these tools.

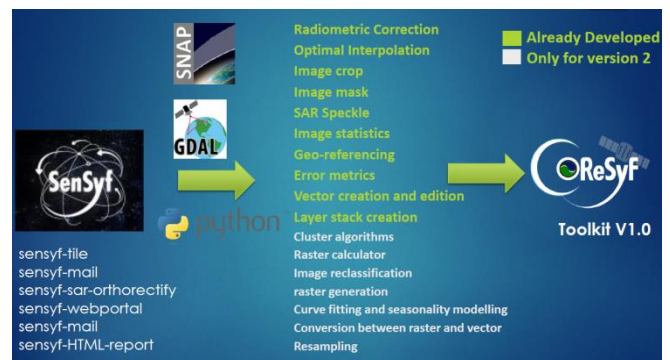


Fig. 4: Co-ReSyf Toolkit

This set of tools, which are the modules for building the processing chains, are just the most basic tools that were identified as crucial to the work to be performed in developing the applications mentioned in Section 4. The idea is for this set of tools to expand with the contributions of the users of the platform in a collaborative way. The usage of open source code in the platform is encouraged, although it is not mandatory. A user can share their tool without exposing the source code. However, should a user wish to share the source code, a project Github repository is available at <https://github.com/ec-coresyf>. The Co-ReSyF toolkit will also be available at this repository.

4. RESEARCH APPLICATIONS

4.1. Core Research Applications

A set of coastal Research Applications (RAs) will be implemented within the Co-ReSyF platform, by the partners of the consortium. The RAs are developed by experts in the respective fields of each RA. These RAs are:

- I. Bathymetry Determination from SAR Images
- II. Determination of bathymetry, benthic habitat classification and water quality from optical sensors
- III. Vessel and oil spill detection
- IV. Time-series processing for hyper-temporal optical data analysis
- V. Ocean coastal altimetry

These core RAs will use platform tools & functionalities to deploy advanced algorithms and methodologies, and will be available for every user. They will also produce a pool of data processing modules, to populate the system and be made available for other users to deploy on other analyses. Any user, of any level of expertise, will be able to access the Co-ReSyF RAs and easily deploy them in their chosen Area Of Interest (AOI). More details on the RAs can be found in [1].

4.2. Applications from Summer School

In mid-2016, the project launched a competitive call for ideas, in order for research students to propose research

subjects using the Co-ReSyF platform. For more details on the competitive call see [2].

From all the applicants, eight selected students were chosen as the successful applicants to perform their research using the platform. The eight subjects selected were:

- I. Predictive habitat modelling for protected cetacean populations in Irish waters using earth observation data
- II. PAR for macroalgae operational modelling
- III. Morphological reconstruction of Tagus ebb-tidal delta in the last 40 years
- IV. Predicting the occurrence of harmful jellyfish species in Irish waters; A tool for aquaculture
- V. Integration of Earth observations in the coastal observatory of Greater Lisbon
- VI. Application for coastline monitoring using Co-ReSyF cloud platform
- VII. ECODE (Extreme COastal Data Evaluation) process EO data: waves, winds and sea level
- VIII. Topo-bathymetry of wave-dominated inlet using Sentinel 2

Their work started with a Summer School organized by the Co-ReSyF project. The Summer School took place from July 12th to 14th in LNEC, Lisbon, Portugal. Scholars were trained to use the Co-ReSyF platform as a key resource for their proposed research project, with continuous technical support on the platform functionality. Eight successful applicants attended the Co-ReSyF Summer School, where they were given access to:

- a) the Co-ReSyF platform,
- b) training material and
- c) technical support to develop their proposed ideas into new coastal applications.

Adding to current Co-ReSyF Research Applications, the development of new ideas allows for these early platform users to demonstrate the capabilities and broader scope of Co-ReSyF to the wider scientific community.

5. LINKING WITH GLOBAL INITIATIVES

Aiming to link with global stakeholders in various areas, such as climate, coastal and EO research, and contribution to Sustainable Development Goals, Co-ReSyF has so far established links with Future Earth Coasts (FEC), EurOcean, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Group on Earth Observations (GEO) Blue Planet, World Meteorological Organisation (WMO) World Climate Research Programme (WCRP), and the NOAA Center for Satellite Applications and Research (STAR). These organisations provide the project with strategic guidance, ensuring the project activities are targeted to achieve maximum societal and scientific benefit.

6. COMMUNICATION AND OUTREACH

Stakeholder-led development is a crucial aspect of Co-ReSyF, iteratively improving and refining the Co-ReSyF platform to address critical barriers the communities face. The team organises various user board and advisory board meetings at different intervals of the project. This ensures the product stays on-track and the platform is as relevant as possible to our target users. So far, the following have taken place:

- First Advisory Board Meeting
- First User board Meeting
- Second Advisory Board Meeting
- Second User board Meeting

All the information is disseminated via our website (www.co-resyf.eu), Twitter (@Co_ReSyF), Facebook (www.facebook.com/coresyf2016) and LinkedIn (www.linkedin.com/groups/8480833).

7. ACKNOWLEDGEMENTS

Co-ReSyF is a 3-year project (2016-2018) funded by the European Union, within the European Union's Horizon 2020 research and innovation programme under grant agreement No 687289.

8. REFERENCES

- [1] Co-ReSyF – Grant Agreement 687289, *Service Definition Document- Research Applications*, European Commission, <http://co-resyf.eu/about-co-resyf/outreach-reports/>, 18/11/2016.
- [2] Co-ReSyF – Grant Agreement 687289, *Competitive Call Opening Announcement and Regulations*, European Commission, <http://co-resyf.eu/about-co-resyf/outreach-reports/>, 18/11/2016.
- [3] ESA Thematic Exploitation Platforms, <https://tep.eo.esa.int/>
- [4] EVER-EST – Grant Agreement 674907, <http://ever-est.eu/>
- [5] EO4WILDLIFE – Grant Agreement 687275, <http://eo4wildlife.eu/>
- [6] EO4ATLANTIC, <https://docs.eo4a.science/overview.html>
- [7] NEXTGEOSS – Grant Agreement 730329, <http://nextgeoss.eu/>

OVERVIEW OF THE ESA ATMOSPHERIC DATA CENTER: EVDC

*Paolo Castracane¹, Angelika Dehn², Paul Kiernan³, Shane Carty³,
Ann Mari Fjaeraa⁴, Thomas Espe⁴, Ian Boyd⁵,
Alastair McKinstry⁶, Conor Delaney⁶, Johannes Hansen⁶*

1-RHEA SYSTEM S.p.A.; 2-ESA/ESRIN; 3-Skytek
4-NILU; 5-BC Scientific Consulting LLC; 6-ICHEC

ABSTRACT

The ESA Atmospheric Validation Data Centre (EVDC) serves as a central, long-term repository for archiving and exchange of correlative data for validation of atmospheric composition products from satellite platforms. EVDC is currently undergoing major upgrades to meet the needs of the upcoming atmospheric composition/dynamic missions in particular for Sentinel-5P and ADM-AEOLUS.

The EVDC builds on the previous ENVISAT Cal/Val database system in operation at NILU since the early 2000s and provides tools for extraction, conversion and archival. The objective of the current ESA funded project (ESA/RFQ/3-14383/15/I-SBo), lead by Skytek [1] with the partnership of NILU [2] and ICHEC [3], is to provide an online information system that supports users in managing and exploiting campaign datasets for Earth Observation missions and applications.

The EVDC will provide access to satellite data subset for specific missions over user-defined areas. Having both Cal/Val data and satellite products in a centralized system will greatly increase the possibility of validating missions over long time series and will improve understanding of the behavior of sensors during the entire mission.

Index Terms— Cal/Val, EVDC, Sentinel-5P, ADM-AEOLUS, GEOMS

1. THE ATMOSPHERIC CAL/VAL DATA

The EVDC contains a large variety of data used for validation of the satellite atmospheric composition products from campaigns, in-situ ground-based measurements, aircraft, balloons and, in general, from a wide range of stations and measurements principles. As an example of EVDC currently available campaign data, in preparation for the Sentinel5-P/TROPOMI data characterization and

validation activities, a dedicated inter-comparison campaign (CINDI-2) for Max-DOAS instruments has been carried out in September 2016 in Cabauw, The Netherlands. The data from this campaign will be used apart from scientific purposes to initialize and characterize a set of Fiducial Reference Measurements for Ground-Based DOAS Air-Quality Observations (FRM4DOAS) [4] network of MAX-DOAS instruments.

The EVDC offers, moreover, the access to other Cal/Val datasets by the mirroring of specific databases (e.g. NDACC [5], ACTRIS [6]) and the provision of the ECMWF data for daily updated global gridded meteorological parameters. The EVDC portal also includes several tools supporting the user in terms of data query, data upload/download, format conversion (GEOMS [7] conversion routines) and for production of ECMWF parameter’s maps.



Figure 1 Examples of EVDC Cal/Val data and Campaigns

Data exchange with the EVDC is regulated by a protocol with the aim to ensure data ownership, to prevent redistribution to third parties and to protect intellectual properties.

2. THE GEOMS FORMAT

The Generic Earth Observation Metadata Standard (GEOMS) is implemented at the EVDC data centre in order to ensure harmonised specification and reporting of data and metadata. GEOMS is developed in collaboration with ESA, NASA and NDACC. In this light all EVDC data should be formatted as netCDF, HDF4 or HDF5 and be compliant to the GEOMS standard, to enhance the usability of correlative data in Cal/Val and ensure an extensive quality control. The EVDC offers numerous conversion tools and specific support for conversion to the GEOMS format. It is possible, however, to collect data with different format.

3. METADATA HARVESTING AND SHARING

EVDC is connected to other data archives through data harvesting technologies. Currently, Atmospheric EO and Cal/Val data are available from multiple sources and data archives across the world. In order to facilitate simpler and faster search methods for the users, EVDC is setting up harvesting methods for sharing metadata between data archives from a number of national and international projects and programs. There is a growing interest in using Cal/Val data, particularly in connection with the new Sentinel missions and other upcoming satellites, as well as in Copernicus and related initiatives. Through metadata sharing, EVDC aims to encourage cooperation between the various data archives, promote open data policy and strengthen collaboration throughout EO disciplines in the best possible way. Metadata sharing leads to: a) data available to more users; b) larger contribution rate in publications; c) proper acknowledgements and more visibility; d) data can be understood and interpreted by any user.

The OAI-PMH technique [8] and "behind-the-scenes" information is provided in detail at EVDC web portal. To register your archive in this initiative and to set up the required protocols, please contact the EVDC team (nadirteam@nilu.no). The database management team will help you getting started and provides front line support for setting up harvester services.

4. THE SATELLITE ELEMENT

Beyond the Cal/Val data, the EVDC will provide access to satellite Level-1 and Level-2 products for specific missions over user-defined areas, this will greatly increase the possibility of understanding the behavior of the on-board instruments over long time series by allowing their validation versus ground-based measurements. With this aim, a specific task of the project is dedicated to the deployment of the infrastructure to download, archive and make available the satellite products for specific missions. The EVDC currently stores and make available Sentinel-3A data for a number of validation locations in France, Ireland, Italy, and Namibia, as a part of a pilot demonstration study.

The target is to include Sentinel-5P and ADM-AEOLUS products. This data store infrastructure, handled by ICHEC, is a Simple Storage Service (S3), implemented using the OpenStack Object Storage (Swift) technology [9].

5. THE NEW EVDC PORTAL

The switch to the new domain <https://evdc.esa.int> occurred during August 2017 and a dedicated newsletter [10] on the release of the new portal was sent out to the EVDC users. The newsletter was distributed by e-mail and made available on the EVDC web pages under the "News and Information" section, as well as published on ESA Earth Online [11] and ESA Sentinel Online [12].

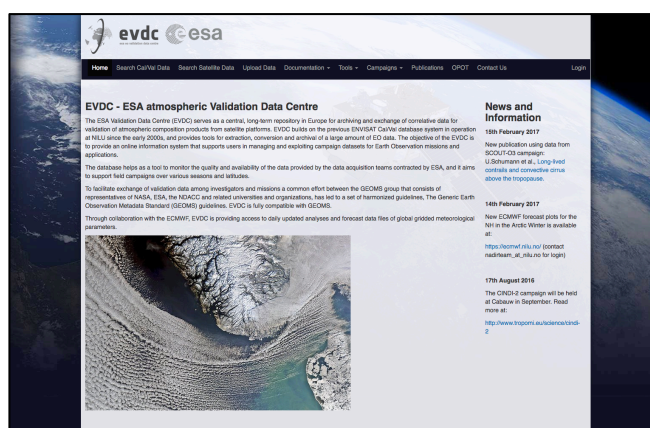


Figure 2 EVDC portal home page



Figure 3 Screenshot of the ESA Earth Online publication

The new EVDC web portal is based on a high scalable Django web application deployed on Amazon Web Service AWS [13]. The main upgrades with respect to the previous portal consist of an improved data access and query performance for the Cal/Val database, the new Search Satellite Data page for accessing the Satellite Archive, and the inclusion of a new feature: the Orbit Predictor and Overpass Tool (OPOT).

5.1. Data Access

The Cal/Val data access in terms of query, upload and download has been strongly improved. The new implementation is based on an Object Relation Mapping (ORM) layer, which abstracts away any database specific details, allowing the focus to remain on the underlying information held within the database. The Satellite products are accessible by an OpenSearch mechanism. The search functionality is provided by means of a searchable index constructed from the metadata of the Satellite files. The index and searching functionality is served by an Open Source search platform called Solr [14]. See below the EVDC panels for CalVal and Satellite data searching.

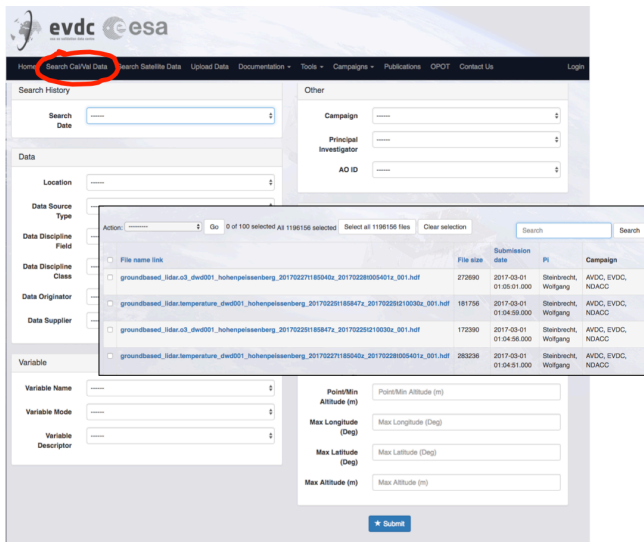


Figure 4 Search CalVal data

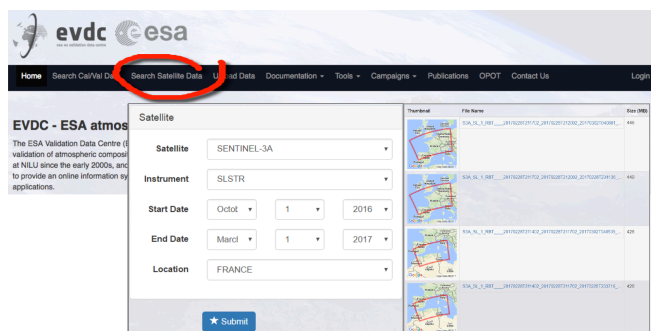


Figure 5 Search Satellite Data

5.2. Orbit Predictor and Overpass Tool

The EVDC portal includes the Orbit Predictor Overpass Tool (OPOT), which takes input for satellite's orbit in the form of TLEs (Two-Line Element set) for specific satellites

and uses the Simplified General Perturbation Model (SGP4) [15] to predict and store their future orbits. Given a location or a Region of Interest (defined as a polygon) the OPOT produces a list of overpasses for that region and satellite for a future time range. Once defined a temporal window, cross-overpass for two satellites is also detectable. Simulated TLEs for future missions (i.e. Sentinel-5P and ADM-AEOLUS) have been included taking into account their orbital characteristics.

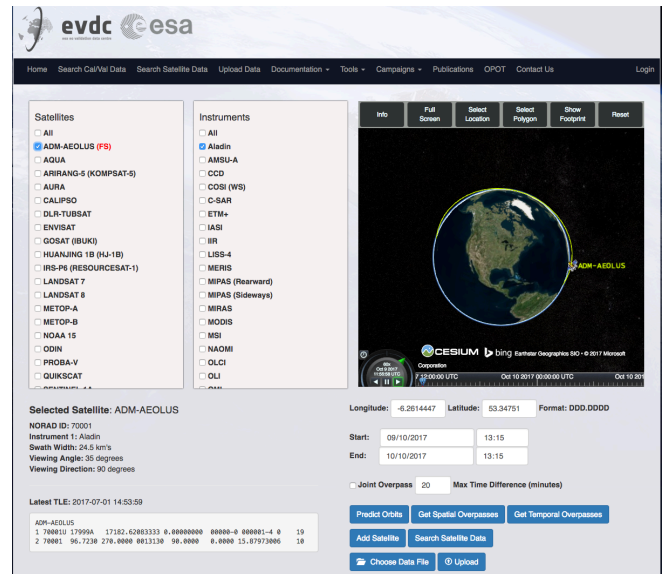


Figure 6 OPOT main panel

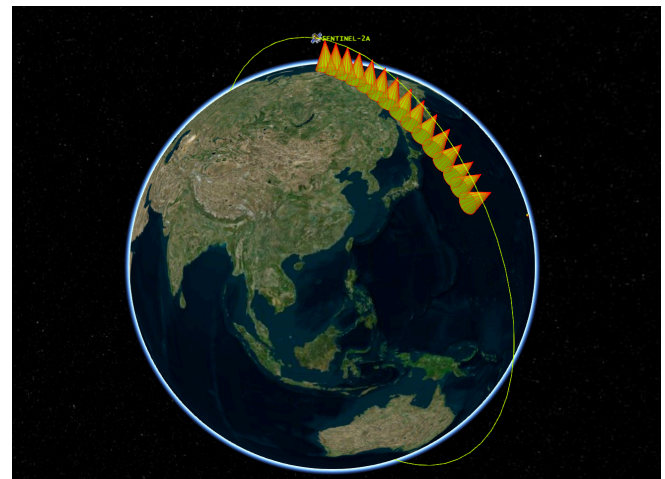


Figure 7 OPOT Example of footprint visualization

5.3. Data Sub-Setting tool

The EVDC Sub-Setting facility, not yet available at the EVDC portal, is currently under development. This facility uses the HARP toolkit [16] as its backbone residing on the EVDC cloud deployed server platform. In its final version it

will be fully integrated into the EVDC site, this integration will give users two options on using the tool: standalone and embedded.

In standalone mode, the sub-setting tool will be accessible from the main menu in a dedicated page, which allows the user to upload a file, perform operations using the HARP command line, and download the resulting file. This will be for users who already have a data file, which they need to convert or filter. The usage of the HARP system as the backbone is totally transparent to the user and the facility provides a custom web based GUI to access the features of the underlying tool.

The user will also be able to use the sub-setting tool from the Satellite Search results page in a more indirect manner (embedded). This functionality is planned to become available starting from the SSP operational phase. Users searching for satellite data, can choose to perform a sub-setting operation on the file instead of downloading the entire file.

By providing a web frontend to the HARP toolkit, the sub-setting facility removes the need for users to install the toolkit on their own machines. It also removes the need for the users' machine to bear the weight of all the computation required when processing large files, which is instead performed by the Sub-Setting facility cloud based server.

6. USER'S INVOLVEMENT

Great importance has been given, to the users involvement. Once the beta version of the portal was implemented (end of Nov. 2016), 25 selected users from different contexts were selected for a period of test of the new functionalities lasting from 4th Dec. 2016 to 20th Jan. 2017. At the end of this testing phase it was requested to fill a questionnaire and provide feedback on the new functionalities of the portal. Responses and comments from users has been analyzed and converted into a list of items, which have been taken into account in the final implementation.

7. CONCLUSION AND ACKNOWLEDGEMENT

The ESA Atmospheric Validation Data Center has a new portal <https://evdc.esa.int> with upgraded functionalities and new tools. The portal offers access to both Cal/Val data and Satellite products for specific missions. The user community is invited to exploit these new features contacting the EVDC team (nadirteam@nilu.no) for any further information.

The evolution of the new EVDC portal has benefitted from the contribution of scientists from various Cal/Val teams and institutes:

- BIRA <http://aeronomie.be>;
- EUMETSAT <https://www.eumetsat.int>;
- FMI <http://en.ilmatietaenlaitos.fi>;
- JPL/NASA <https://www.jpl.nasa.gov>;
- KNMI <http://www.knmi.nl>;
- MPI <https://www.bgcjena.mpg.de>;
- and RHEA <https://www.rheagroup.com>;

who have been involved in the test phase.

8. REFERENCES

- [1] Skytek <http://www.skytek.com/>; info@skytek.com
- [2] NILU Norwegian Institute for Air Research. <http://www.nilu.no/>; nadirteam@nilu.no.
- [3] ICHEC Irish Centre for High-End Computing. <https://www.ichec.ie/>
- [4] FRM4DOAS <https://earth.esa.int/web/sppa/activities/frm>
- [5] NDACC Network for the Detection of Atmospheric Composition Change <http://www.ndsc.ncep.noaa.gov/>
- [6] ACTRIS Aerosols, Clouds, and Trace gases Research InfraStructure network <http://www.actris.eu/>
- [7] GEOMS host website at NASA/GSFC: <https://avdc.gsfc.nasa.gov/index.php?site=1178067684>
- [8] Open Archive Initiative Protocol for Metadata Harvesting <https://www.openarchives.org/pmh/>
- [9] Swift technology documentation: <https://docs.openstack.org/developer/swift/>
- [10] EVDC Newsletter: <http://ext.mnm.as/v/C86C9692-C22A-4632-8693-A0755A371BA9>
- [11] ESA Earth Online: <https://earth.esa.int/web/guest/missions/user-services-news>
- [12] ESA Sentinel Online: <https://sentinels.copernicus.eu/web/sentinel/news/-/article/new-portal-for-the-esa-atmospheric-validation-data-center-evdc->
- [13] Django web framework documentation : <https://www.djangoproject.com/>
- [14] Solr documentations. <http://lucene.apache.org/solr/>
- [15] Models for Propagation of NORAD Element Sets. F. R. Hoots R. L. Roehrich 1980. <https://www.celestrak.com/NORAD/documentation/spacetrk.pdf>
- [16] HARP documentation: <https://cdn.rawgit.com/stcorp/harp/master/doc/html/index.html>

FOOD SECURITY – THEMATIC EXPLOITATION PLATFORM BIG DATA FOR SUSTAINABLE FOOD PRODUCTION

Heike Bach¹, Silke Migdall¹, Markus Muerth¹, Phillip Harwood², Andrea Colapicchioni², Antonio Cuomo², Sven Gilliams³, Erwin Goor³, Tom Van Roey³, Andy Dean⁴, Jason Suwala⁴, Antonio Romeo⁵, Esther Amler⁵, Philippe Mougnaud⁵, Espen Volden⁵

¹VISTA GmbH, Gabelsbergerstr, 51, 80333 Munich, Germany, email: bach@vista-geo.de

² CGI Italia S.r.l, Via Enrico Fermi 62, 00044 Frascati (RM), Italy

³VITO, Boeretang 200, B-2400 Mol, Belgium

⁴ Hatfield Consultants, 200, 850 - Harbourside Drive, North Vancouver, BC, Canada

⁵ESA ESRIN, Largo Galileo Galilei 1, 00044 Frascati RM, Italy

ABSTRACT

The innovative Food Security Thematic Exploitation Platform of the European Space Agency fosters smart, data-intensive agricultural and aquacultural applications in the scientific, private and public domain with immediate access to Earth Observation (EO) data, other related data, and data processing tools. Additionally, tailored EO-based services will be implemented on the platform to support the Food Security community.

Index Terms— Food Security, Thematic Exploitation Platform, cloud processing, biophysical parameters, sustainable food production

1. INTRODUCTION

The last years have seen a rapid growth in EO data availability. The Sentinel satellite constellation developed and operated by ESA within the European Commission led EO program Copernicus, routinely monitors our environment globally, delivering an unprecedented amount of open data. This data, in combination with in-situ networks and models provides unique possibilities of analysis, both on the local scale (e.g. providing crop status for site-specific applications) as well as on regional and country-wide scales (e.g. for insurance services in Africa).

With the growing data volume also comes the challenge of achieving full potential in terms of data exploitation. In this context, ESA started the “Thematic Exploitation Platforms” initiative [1]. The Food Security-TEP (FS-TEP) is the newest in a range of TEPs. It aims at providing a “one stop platform” for the extraction of information from EO data for data-intensive services in the Food Security sector mainly in Europe & Africa, allowing both access to EO data and direct processing of these on the platform. This aligns with the paradigm shift of bringing the users to the data to enable the processing of large datasets.

The TEP builds on a large and heterogeneous user community, spanning from application developers in

agriculture to aquaculture, from small-scale farmers to agricultural industry, from public science to the finance and insurance sectors, from local and national administration to international agencies.

2. CONCEPT

The FS-TEP backbone (Fig.2.1) provides Earth Observation data and products as well as tools and methods that are relevant for the Food Security user community. To maximize the agility and functionality of the platform, as far as possible existing infrastructure and software will be integrated in a federation of platforms. Starting with the infrastructure and data access as provided by an existing DaaS and IaaS service like CloudFerro (the Polish testbed) or Code-De (the German Copernicus platform of DLR). Software from the Forestry-TEP is integrated as well as the platform and information services of the Proba-V MEP. Available datasets are integrated as InaaS. This is all input to the platform and software of the FS-TEP that leads to the information service (FS-TEP InaaS).

On top of the platform backbone, the User Front Office Services, delivered through the Graphical User Interface (GUI) are classified in 3 components with varying offerings and concepts for the different user groups:

- An interface for expert users will be the entry point to the comprehensive functionality of the platform including fast and easy access to satellite and ancillary data as well as the tools to explore, analyse, and process these datasets.
- A mobile version of the platform will allow on-site visualization of EO information products and their time series via smartphone and is designed for easy, intuitive use.
- A customized version will provide additional, user-adapted information and interfaces for the monitoring and management of fish and agriculture farms as well as confidential data management in a secure environment.

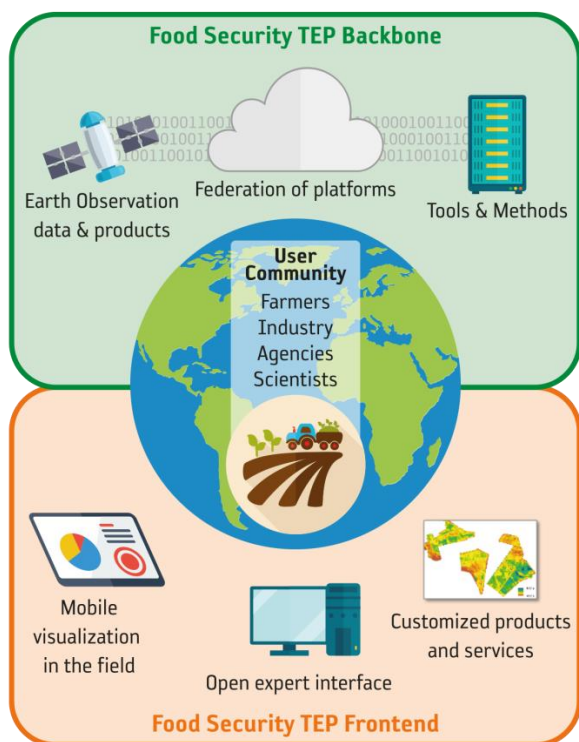


Fig. 2.1: Concept of the FS-TEP

The following data and tools are foreseen to be available in the “Essential” part of the TEP after the first year:

- Satellite data will include Sentinel-1,2,3, Landsat 8 and Landsat legacy ESA as well as the full Proba-V data archive. Pre-processed atmospherically corrected Sentinel-2 data as well as a set of biophysical plant parameters will be available for an initial service area (Germany, Belgium, the Netherlands, selected parts of Zambia)
- As ancillary data terrain maps, soil maps and some meteorological data will be available.
- As tool boxes the SNAP Toolbox, QGIS and ORFEO/Monteverdi will be available.

The FS-TEP is developed under an agile development approach, so changes to the service extent are possible. The development takes place in two phases, with the first phase covering the main functionalities as well as the first service pilot, and the second phase covering additional functionalities and the following two service pilots (see section 4 and figure 2.2).

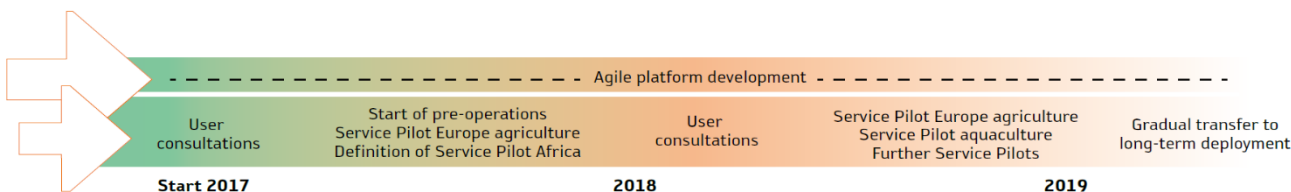


Fig. 2.2: Timeline for the FS-TEP development

3. USER COMMUNITY NEEDS

In June 2017, two dedicated user workshops were conducted in order to assess gaps & needs in EO information services and to involve the community in the service definition.

The workshop participants came from all relevant agricultural sectors: public, farming, finance and Earth Observation. Many of them are participants in the “Partnership for Growth and Sustainability”, a consortium of experts in the field of Food Security which helped ESA defining the requirements for the platform and enables the team to continually develop the platform in accordance with the user community’s needs. [2].

Summarizing the outcomes quickly, the following statements can be made:

- All user groups are interested in using the FS-TEP.
- Research institutes are most interested in bringing their own algorithms to the TEP, while the private industry likes the option of customized services best, where they can buy services from expert service providers.
- Almost all user groups are interested in having a mobile service that allows the viewing of certain plant parameters on the field with a mobile device.

As expected, with a user group as diverse as the one for the FS-TEP, when it comes to the details, there are many different wishes and requirements, concerning both functionalities and available data sets. However the users agree on, that having ancillary data of good quality – e.g. topographical data, soil maps, meteorological data, and administrative data - on the same platform as the EO data, is a feature which has so far been missing. A Keyword here is “good quality”: Rather than having the largest amount of possible options, most users would prefer access to high quality data sets and services available on the FS-TEP.

4. SERVICE PILOTS

In order to demonstrate the platform’s ability to support agriculture and aquaculture with tailored EO based information services, service pilots will be implemented. Altogether there will be three service pilots, which will be designed to showcase the functionality of the FS-TEP for agriculture and aquaculture in both Europe and Africa and especially under the aspect of big data handling.

The three service pilots will target different aspects of Food Security that can be supported by EO data, with the first service focusing on the sustainable increase of agricultural production efficiency. This is in line with the UN's Sustainable Development Goal of "Zero Hunger" [x], for which more food per area of current farmland will need to be produced in a sustainable way. Within the Copernicus satellite constellation, especially Sentinel-2 provides the means to establish a comprehensive information chain from the satellite to the farmer.

By the application of similar EO based and information driven solutions, the first service pilot is not restricted to a European or African case, but targets to address both regions with their varying challenges and different agricultural production schemes.

In Africa, the yield of cereals per hectare has only slightly increased from 1960 to 2014. Europe reached 6.8 t/ha on average in cereal production, whereas Africa only achieved 1.6 t/ha (year 2014) [3]. Yield gaps as well as the demand for a sustainable intensification in Africa are high. Considering favorable climatic conditions for cropping on African farmland, there is significant potential for increased efficiencies and for closing the large current yield gap.

One reason for higher yields in Europe is the higher amount of resource input. The use of fertilizer in Europe is currently more than 6 times higher than in Africa. It is expected that in African agricultural systems fertilizer and plant protection products input will increase. Advanced management techniques can be used to avoid European mistakes and to not intensify agriculture above a sustainability level. These techniques can assist in the reduction of fertilizer loading of the groundwater caused by nitrate leaching, and in the preservation of soil fertility, which is endangered by erosion and salinization through inappropriate irrigation techniques. Hence, smart farming can provide sets of ecologically and economically meaningful measures to improve productivity. That increase in productivity can also have positive impacts on water issues. Measurements showed that with higher yields the water productivity increases [4] and "more crop per drop" can be achieved.

Since the first service pilot is set to demonstrate the capabilities of the FS-TEP already within the first year of its development, the pilot applications are based on existing EO-based agricultural services, namely VISTA's TalkingFields [5] and VITO's WatchItGrow [6]. These smart farming services go a step beyond the original "Precision Farming", which is based mainly on farming technology to e.g. allow for auto-steering of tractors and harvesters. The focus of smart farming shifts towards a more rounded, holistic approach - going from "highest spatial precision" to "smartest treatment". Thus, typical issues of smart farming are e.g. how much fertilizer is best applied when and where in the field or which plant protection resources are optimal for crop development at each location in the field. Although these smart farming techniques are not yet common in Africa and

still under development in Europe and North America, they can and should be adapted to the African case.

Figures 4.1 and 4.2 show an example of biophysical plant parameters derived from Sentinel-2 that can be used as input for fertilization advice. The leaf area gives an indication of the amount of biomass on the field while the chlorophyll content is correlated with the nitrogen content in the plants. Together, these two variables can be used to calculate the nitrogen demand of the crop.



Fig. 4.1: Leaf area of maize derived from Sentinel-2 (Source: VISTA)



Fig. 4.2: Leaf chlorophyll content of maize derived from Sentinel-2 (Source: VISTA)

Biophysical plant parameters like this will be available pre-calculated on the FS-TEP so that they can both be used in the mobile version to directly visualize the current growth on single fields, and also as input to different other services. During the demonstration phase, leaf area, chlorophyll content, fraction of absorbed photosynthetically active radiation and fractional cover will be provided on the platform. Service regions for the derivation of biophysical plant parameters will be on the country-scale. The first service pilot will focus on Germany, the Netherlands, Belgium and parts of Zambia, with the service region to be extended in the second year for the following two service pilots. These will focus on Africa and aim at supporting financial and insurance schemes for small scale farmers as well as at providing remote sensing support for aquaculture in Africa.

5. OUTLOOK

The FS-TEP is currently in its first year of development. The first service pilot will be pre-operational in the beginning of 2018, while the later services will be defined in detail in a further user workshop in spring 2018 and will be pre-operational towards the end of 2018. News and achievements, further information as well as a user forum can be found at <http://foodsecurity-tep.eo.esa.int>.

6. ACKNOWLEDGEMENTS

The project team developing the FS-TEP and implementing pilot services during a 30 months period (started in April 2017) is led by Vista GmbH, Germany, supported by CGI Italy, VITO, Belgium, and Hatfield Consultants, Canada. It is funded by ESA under contract number 4000120074/17/I-EF.

REFERENCES

[x] UN, 2017: Sustainable Development Goals, <http://www.un.org/sustainabledevelopment/sustainable-development-goals/>.

[1] ESA, 2017: Thematic Exploitation Platform, <https://tep.eo.esa.int/>.

[2] EO4Food, 2017: Earth Observation Needs and Opportunities to Support Sustainable Agriculture and Development, <http://eo4food.org/>.

[3] World Bank (2014): Agriculture & Rural Development, <http://data.worldbank.org/topic/agriculture-and-rural-development>.

[4] Zwart SJ, Bastiaansen WGM. (2004): Review of measured crop water productivity values for irrigated wheat, rice, cotton and maize. *Agric. Water Management* 69:115-133.

[5] Vista , 2017: Talking Fields, www.talkingfields.de.

[6] Vito, 2017: Watch it Grow, <https://watchitgrow.be/en>.

EO4WILDLIFE: A CLOUD PLATFORM TO EXPLOIT SATELLITE DATA FOR ANIMAL PROTECTION

Fabien Castel¹, Gianluca Correndo², Alan F. Rees³

¹Atos Integration, Toulouse France

² University of Southampton, IT Innovation Centre, Southampton United Kingdom

³ University of Exeter, Penryn, UK

ABSTRACT

EO4wildlife brings large number of multidisciplinary scientists such as marine biologists, ecologists and ornithologists to collaborate closely together while using European Sentinel Copernicus earth observation data more efficiently on a platform available over Internet dedicated to environment study and animal protection [1]. A comprehensive set of processing services and data connectors are available on the platform providing to scientists powerful tools to build innovative animal protection applications.

Index Terms— Copernicus, Platform as a service, big-data, cloud computing, earth observation data, data analytics, animal protection

1. INTRODUCTION

All the new sets of data provided by Copernicus satellites open up the way for hundreds of innovative scenarios to combine animal tracking data with remotely-sensed earth observation data. In order to reach such important capabilities, an open service platform and interoperable toolbox supported by a scalable cloud infrastructure are being designed and implemented. It offers high level data processing services. The platform front end will offer dedicated services that will enable scientists to connect with several animal tracking databases, access large data collections from Copernicus satellites, sample relevant environmental indicators, and finally run environmental models and simulations using these big data sources in a scalable processing environment.

The research is leading to the development of web-enabled service compliant to OGC2 (Open Geospatial Consortium) enabling data interoperability in geospatial data access and processing services.

2. PROBLEMATIC

Scientists can use the huge Copernicus datasets for various purposes. Mainly, they aim at identifying the key environmental factors that drive the distributions of animals. By building predictive models the goal is to improve

management and decision-making about animal protection. Model results are extrapolated, in line with various climate change scenarios, to determine likely future population distributions and aid understanding of how environmental conditions may alter phenology and demographic processes. Exploiting these rich datasets is a challenge for scientists. A wide diversity of products are available through different systems, platforms and interfaces, but this profusion of options can be overwhelming and scientists do not always have the technical capabilities to access these sources and to process the downloaded data. That's why the EO4wildlife platform aims at providing a quick and easy access to a comprehensive set of EO datasets, as well as a toolbox of services for data filtering, processing and visualization.

3. CLOUD APPLICATION

Generally, cloud application stands for applications deployed over Internet with flexible pay-as-you-use infrastructure, "big-data" storage and scalable web services. The EO4wildlife platform perfectly fit with this description. In particular, several layers can be distinguished when dealing with applications on the cloud. The Infrastructure as a Service layer – IaaS – provides flexibility and efficiency in order to easily scale out processing and storage capabilities. The Platform as a Service layer – PaaS – deals with applicative components deployment, resource management and security issues. The Software as a Service layer – SaaS – allows user to access the required data and to configure and run the service available on the platform.

There are many benefits with such an approach. Besides the economical and practical aspects, there is a strong incentive for sharing. On the platform, everyone can be both a producer and a consumer, and discover new opportunities within the community. EO4wildlife offers a catalogue of resources and added value services that is continuously enriched as new members join and contribute to the ecosystem.

4. DATA ACCESS

4.1. Tracking Data

EO4wildlife aims at being connected to existing platforms where scientists host their data. Initially the Seabird (<http://seabirdtracking.org/>) and the Seaturtle (<http://seaturtle.org/>) tracking databases are targeted.

The Seabird Tracking Database aggregates data from more than 150 contributors to provide the largest collection of seabird tracking data in existence. In total, the Seabird Tracking Database holds information for 114 species in more than 11 million locations, corresponding to more than 20 thousand tracks. It serves as a central store for seabird tracking data from around the world and aims to help further seabird conservation work and support the tracking community.

The seaturtle.org platform hosts Argos tracking data for all seven species of sea turtle and around 70 other animal species that include cetaceans, pinnipeds, elasmobranchs and birds located around the globe. The platform hosts data for over 1000 tracking projects and has amassed over 14 million data points.

4.2. Environmental Data

The EO4wildlife platform provides connectors to access data from various data sources, such as the CMEMS [2] catalogue, indexing hundreds of ocean-related EO products, or the AVISO catalogue for the altimetry domain. An internal EO4wildlife data catalogue is maintained to aggregate products from all these external sources.

Environmental datasets generally are voluminous. The Figure 1 is an example of a dataset that is used on the platform. This dataset provides data on sea surface temperature and sea ice concentration during a 10 year period from 2007 to 2017 for a global geographic coverage. The overall size for the dataset is near 0.5 Terabytes. The platform aims at targeting tens of datasets similar to this one.

5. PROCESSING SERVICES

The platform hosts a series of basic data analytics services to enable scientists to implement analytic workflows [3]. These services can be divided in three main categories.

5.1. Pre-processing and Aggregation

The first one is **data pre-processing and aggregation**. It includes the pre-processing, cleaning and aggregation of the data prior to the analytical step. The pre-processing of geospatial data sets is an important step when dealing with potentially imprecise information such as animal positions. These services allow to recognize and to eliminate all data elements which are clearly unrealistic considering the knowledge of the domain. (E.g. the animal is not capable of

travelling at such velocities). Moreover, the pre-processing services allow filling missing data values and accommodate different data grids by interpolating values which were not directly collected and represented in the data sets. Aggregation services reconcile data represented with different spatial or temporal resolutions providing functionalities to sample environmental observations and aggregate them in the right granularity to fuel niche modelling algorithms. In this category are also included services to process animal tracks, to provide grouping of tracks in trips or gridding a number of tracks to study the population distribution.

Global Ocean OSTIA Sea Surface Temperature and Sea Ice Analysis			
Variables	size/day (MB)	size/year (GB)	total size (GB)
analyzed_sst	50	18,25	196,55
sea_ice_fraction	25	9,125	98,275
analysis_error	50	18,25	196,55
Whole product	125	45,625	491,375

Figure 1: EO Dataset volume

5.2. Data Mining

The second category is data mining and contains services processing animal tracks and satellite marine observations in order to model animals' use of space and correlate this information with available environmental observations. This category is further subdivided in two sub-categories of services: animal tracks based services and statistical environmental services.

Animal tracks based services analyse the tracks alone in order to estimate the animals' home range and the foraging grounds.

Statistical environmental services assess the statistical relevance of environmental observations in modelling animals' presence and implement environmental niche models (ENM) to understand the marine species' habitats.

Data mining techniques have been implemented in different scenarios to model the preferred animals' habitat following a literature review for each marine species. At the moment these modelling techniques include:

- Environmental Envelope Model (EEM)
- Generalised Additive Model (GAM)
- Generalised Linear Model (GLM)
- Boosted Regression Trees (BRT)
- Random Forests (RF)

These services support scientists in modeling marine animal niches by means of environmental observations which can then be used in creating projections of animal presences in the future (see Figure 2 for an example of niche model produced using RF).

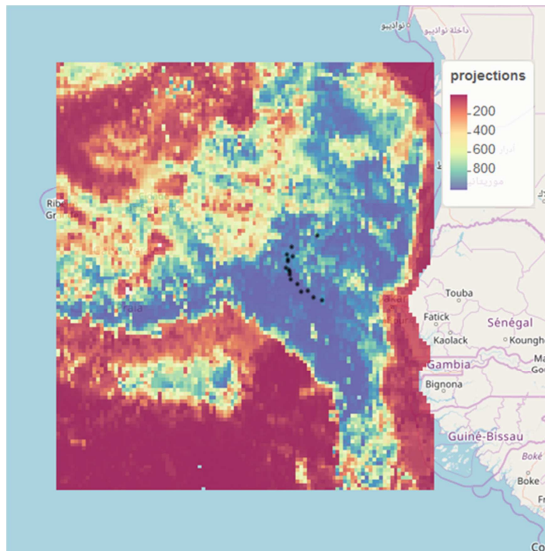


Figure 2: Marine turtle habitat modelled using RF near Cape Verde (August 2009)

5.3. High Level Fusion Services

The last category contains **high level fusion services**. These services make use of multiple data sources to better estimate animals' position, behavior and modelling animals' habitats. This category includes the Track & Loc service which enables the estimation of submarine trajectories for animals equipped with pop-up or archival tags [6].

6. IMPLEMENTATION

6.1. Architecture Overview

The platform is composed of several functional components. An internal data catalogue aggregates georeferenced products metadata from various external sources. An ingestion component allows retrieving this data on-demand for exploitation by the platform services. The service manager component allows developers to manage the lifecycle and the execution of their services. At the end of the chain, EO4wildlife makes available built-in visualization features for standard geographic data (OGC WMS/WFS standards) produced by the services.

The service management mechanism is built on the containerization concept (i.e. Docker [4]). By encapsulating each service into an independent and self-sufficient container, the platform ensures total freedom for the service developers (preventing language, framework or libraries constraints) and an easy portability on the cloud. An

orchestration technology (i.e. Kubernetes [5]) is used to manage container life cycle so that the underlying infrastructure becomes totally transparent. This technical architecture based on standards aims at creating a collaboration space for scientists in term of data and service sharing

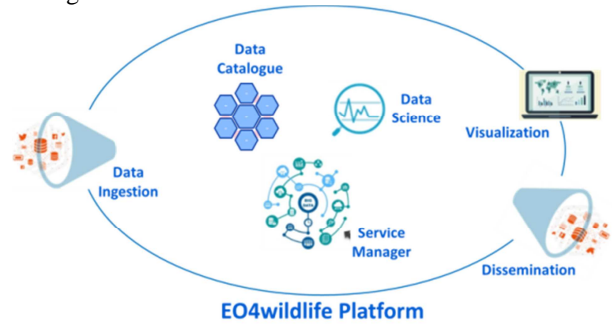


Figure 3: Functional view

6.2. Data Management

The ambition of EO4wildlife is to grant scientists easy access to tens of earth observation datasets, thus dealing with terabytes of data (see 4.2). Instead of hosting permanently all the data to the platform, a repository is set up and acts as a local cache from the remote data warehouse. Data is day by day downloaded following platform user's requests and temporarily stored on the local cache. This cache is common to all users. When providing data to a service, the platform deals with the merging of these daily files in order to have one global file corresponding to what was requested for the service execution. Finally, a periodical purge mechanism ensures that the cache size does not overreach the local disk size.

6.3. Interoperability

Integrating smoothly the EO4wildlife platform into the existing ecosystems of animal monitoring applications is a key element for scientists. EO4wildlife was designed to complete existing systems. For this purpose, it provides convenient interface for external application to upload data into the platform. Indeed the internal module in charge of the file management on the server side exposes a REST web service enabling any external application (Seabird and Seaturtle for instance) to upload data to the platform. Data exchange between different platforms can be challenging because every tracking database has its own format, generally csv-like text files. To address this issue, the EO4wildlife project defined an XML format specific to the animal tracking domain. Any tracking data can be converted to this format using a configurable CSV to XML converter. Configurations for the Seabird and Seaturtle file format are already deployed on the platform. All the services offered by the platform should use this standard format. Besides an improvement on services coherence and

versatility, the standard self-described format has other advantages. Date information becomes much more exploitable, for instance for automatic animation over time of the tracking points and automatic extraction of spatial coordinates is possible when fusing tracking and earth observation data.

6.4. Data Visualization

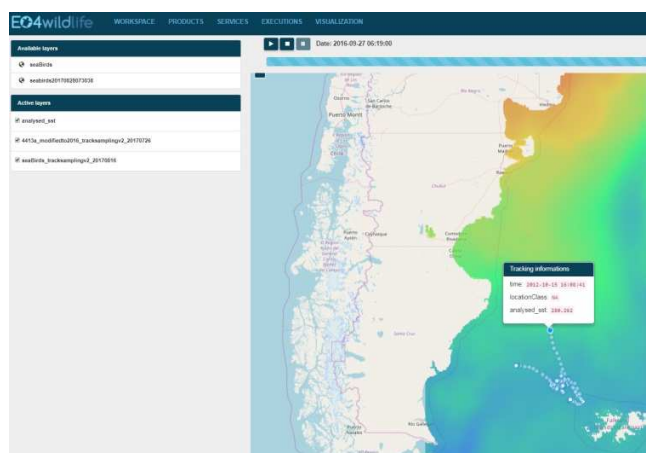


Figure 4: Visualization panel screenshot

The platform provides a visualization panel (see Figure 3) for tracking and earth observation data. This feature uses the OGC web standards for geographic data representation and a frontend panel based on multi layers visualization. Tracking data are handled as WFS [7] layers and earth observation product as WMS [8] layers. All the data used as inputs and/or outputs of processing services can be automatically transformed into displayable layers that can be overlaid. All the layers holding a time dimension can also be animated over time, which allows the highlighting of the key environmental factors on animal migration.

7. CONCLUSION & NEXT STEPS

This paper presents the services that the EO4wildlife platform can offer to the scientific community in order to develop innovative use cases based on the new generation of earth observation data. The developments of the platform will continue until 2018, developing new features following the needs of the project scientific partners and securing what exists to prepare the opening to a larger public.

8. ACKNOWLEDGEMENT

This work is partly funded by the European Commission under H2020 Grant Agreement number: 687275.

9. REFERENCES

- [1] Zoheir Sabeur, Gianluca Correndo, Galina Veres, Banafshe Arbab-Zavar, Jose Lorenzo, Tarek Habib, Anne Haugommard, Fanny Martin, Jean-Michel Zigna, and Garance Weller. 2017. EO Big Data Connectors and Analytics for Understanding the Effects of Climate Change on Migratory Trends of Marine Wildlife. Springer International Publishing.
- [2] CMEMS (Copernicus Marine Environment monitoring service) public website : <http://marine.copernicus.eu/>
- [3] Z. Sabeur, G. Correndo, G. Veres, B. Arbab-Zavar, G. Ivall T. Neumann, F. Castel, J-M. Zigna, J. Lorenzo. (2017) EO Big Data Analytics for the Discovery of New Trends of Marine Species Habitats in a Changing Global Climate. 2017 Conference on Big Data from Space (BiDS'17)
- [4] R. Peinl, F. Holzschuher, F. Pfitzer, Docker cluster management for the cloud –survey results and own solution, J. Grid Comput. (2016) 1–18.
- [5] D. Bernstein, Containers and cloud: from LXC to Docker to Kubernetes, IEEE Cloud Comput. 1 (3) (2014) 81–84.
- [6] Royer F, Lutcavage M (2009) Positioning pelagic fish from sunrise and sunset times: error assessment and improvement through constrained, robust modeling. In: Neilson JD, Smith S, Royer F, Paul SD, Porter JM, Lutcavage M (eds) Tagging and tracking of marine animals with electronic devices. Springer, Amsterdam, p 323–341
- [7] OGC Web Feature Service standard format description <http://www.opengeospatial.org/standards/wfs>
- [8] OGC Web Map Service standard format description <http://www.opengeospatial.org/standards/wms>

APPLICATION OF EARTH OBSERVATION TO A UGANDAN DROUGHT AND FLOOD MITIGATION SERVICE

Samantha Lavender¹, Paul Healy², Ian Robinson³, Regina Lally⁴, Stephanie Ties⁵, Darren Lumbroso⁶, Elizabeth Valone⁷, George Gibson⁸, Lucrezia Tincani⁹, Caroline Chambers¹, Chris Doyle¹, Andrew Lavender¹, Alexa Williams¹, John Auburn², Elma Jenkins², Arnaud Le Carvenec², Miguel Morgado², Simon Reid², Luca Innocente³, Lisa Osborne³, Heather Pitcher³, Sebastian Clarke⁵, Jamie Williams⁵, Gina Tsarouchi⁶, Jimmy Okori⁷, Mark Harrison⁸ and Richard Jones⁸.

¹Pixalytics Ltd, Plymouth, UK; ²RHEA Group, Harwell, UK; ³AA International & AgriTechTalk International, Aberystwyth, UK & Kampala, Uganda; ⁴Databasix, Harwell, UK; ⁵Environment Systems Ltd, Aberystwyth, UK. ⁶HR Wallingford Ltd, Wallingford, UK; ⁷Mercy Corps, Kampala, Uganda; ⁸Met Office, Exeter, UK; ⁹Oxford Policy Management Ltd, Oxford, UK;

ABSTRACT

The Ugandan Drought and Flood Mitigation Service aims to give practical information to help local communities respond to the effects of extreme weather events, across different timescales (ranging from synoptic to end-of-century climate change). It is being funded as part of the UK Space Agency's International Partnership Programme, and is using satellite Earth observations alongside meteorological and hydrological modelling, and ground-based data within an innovative platform. Free-to-access input Earth observation products, from sources such as the Copernicus missions, are the basis for the onward development of information about flood and drought conditions alongside crops and their growing conditions, with the modelling activities allowing for future predictions to be made. The project is within its first year, of a 4 year programme, and is currently focused on the platform development alongside bringing together the key input data streams and engaging with the community in Uganda to ensure what's developed is fit for purpose.

Index Terms— Agriculture, Copernicus, Data Cube, Drought, Earth Observation, Exploitation Platforms, Floods, Modelling.

1. INTRODUCTION

Uganda is a landlocked country of just over 240 000 square kilometres. Agriculture is a key element of the country's economy, being responsible for 23% of gross domestic product in 2011 and almost half the country's exports the following year [1]. According to the Food & Agriculture Organisation (FAO) of the United Nations, 80% of the population relies on farming for its livelihood [2].

Uganda has an equatorial climate, with regional variations, although recent recurrent dry spells have impacted on crop and livestock productivity [3]. The level of awareness of the effects of climate change is high among African States.

Investment in institutions, structures and policy building is also high, with noticeable international support. Monitoring systems that reach beyond abstract data collection for long term analysis to provide locally relevant real-time advice would provide significant benefit to local populations and international partners generally.

A consortium led by the RHEA Group, working with the Ugandan Ministry of Water and Environment and local Non-Governmental Organizations (NGOs, AgriTechTalk Uganda and Mercy Corps) is developing a Drought and Flood Mitigation Service (DFMS). This aims to give practical information to help local communities respond to the effects of climate change, with forecasts throughout the growing seasons to enable them to take actions to maximise their crop yield. It is funded as part of the UK Space Agency's International Partnership Programme, with 21 projects chosen to provide solutions to local issues in countries across Africa, Asia, Central and South America. The project is within its first year, of a 4 year programme.

The project should benefit local communities by:

- Improving the ability to forecast and mitigate droughts and floods on a local actionable scale.
- Allowing NGOs to target resources saving time, money and lives.
- Allowing farmers to improve their lives and better protect their livestock and crops.

2. METHODOLOGY

2.1. Why a Space-Based Solution?

In-country investigations showed that space-based data and derived products are not used operationally or systematically in Uganda at present. Although the Ugandan government is aware of what various agencies are saying about climate on a regional level, but the approach is

scattered. Therefore, in Uganda, little use is made of Earth Observation (EO)-based products to help to forecast future extreme weather events.

EO can form a powerful base for unifying information from a variety of ground and community resources, whilst establishing skills and technology for processing satellite datasets which may be applied in a wider geospatial context. In data-scarce countries like Uganda, the use of space technologies can make it possible to enhance on-the-ground collection of data, especially in the remotest areas of the country. The use of EO is therefore essential for the success of this project.

2.2. Approach to Monitoring Water Availability

In terms of water monitoring, there is a real lack of ground-based hydro-meteorological and related data. A limited number of climate and river flow gauging stations exist, leading to problems with data density, making it challenging for water agencies to make meaningful assessments of water resources and to predict floods and droughts. According to the Ugandan Ministry of Water and Environment at present there is no available operational flood forecasting system in the country [4].

The platform (see Fig. 1) is designed to assimilate heterogeneous data sources ranging from satellite and modelling data, as well as community/mobile sources which will be used for both developing products related to drought and flood forecasting such as the prediction of precipitation, soil moisture and vegetation indices. It is being built as a set of Application Programming Interfaces (APIs) and Cloud technologies are used for both data storage and processing.

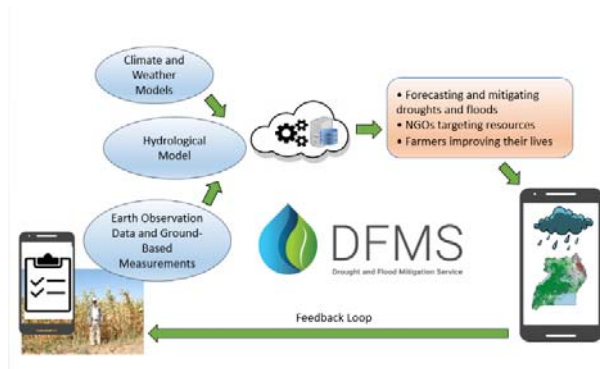


Fig. 1. DFMS platform

Satellite data is stored in Amazon Simple Storage Service (S3) buckets or the DFMS Open Data Cube (ODC) platform, an open source software toolset based on the Committee on Earth Observation Satellites (CEOS) Data Cube [5] for creating a pixel-based time-series of multiple datasets ready for analysis. The justification being that the use of Analysis Ready Data (ARD) will drastically reduce the burden of processing [6], with the Data Cube enabling efficient storage

of large quantities of EO data that will simplify both the access to and use of space-based data.

A hydrological model is used for groundwater and surface water forecasting, alongside the output of climate and weather models that will both drive the hydrological model and be combined with the EO data through statistical modelling, supported by ground-truth yield data collected by Ugandan partners in situ using tried and tested Pictorial Evaluation Tool (PET) methodologies [7]. Microwave and optical EO data are sourced primarily from free-to-access mission datasets such as those coming from the Copernicus Sentinel (1, 2 and 3) satellites, Advanced Scatterometer (ASCAT) on METOP-A, Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) on Landsat-8, Moderate Resolution Imaging Spectroradiometer (MODIS) on both the Aqua and Terra platforms, MIRAS (Microwave Imaging Radiometer using Aperture Synthesis) on the Soil Moisture Ocean Salinity (SMOS) mission, Soil Moisture Active Passive (SMAP) mission and the Visible Infrared Imager Radiometer Suite (VIIRS) on the Suomi National Polar-orbiting Partnership (Suomi-NPP) mission. These products are then converted into ARD and made available through a geoportal alongside the extraction of information for the Ugandan government, especially by the Ministry of Water and Environment and the Ugandan National Meteorological Authority (UNMA), and NGO usage. The output products will provide information about flood and drought conditions alongside crops and their growing conditions.

2.3. DFMS Platform Implementation

The DFMS platform has been implemented with a microservice architecture. Two orchestrators, the Long Lived Service Manager (LLSM) and Ephemeral Service Manager (ESM), respectively manage the long-lived services (which nominally remain active indefinitely) and the ephemeral services (which are deployed and shut down as needed). Fig. 2 shows the high-level DFMS platform architecture. All communication between microservices takes place via the Messaging Service.



Fig. 2. Architecture Overview

The Data Library, where the DFMS ODC lives, acts as an interface to the Services database (data cubes and S3

buckets). The platform has three databases: the Long Lived Services Repository, the Ephemeral Services Repository, and the Logging Database. The Long Lived Services Repository interfaces with the LLSM, while the Ephemeral Services Repository interfaces with ESM.

For each DFMS Service, the processing component produces DFMS model inputs. The data exchange component stores the ingested inputs as ODCs or buckets on the cloud. The ingestion component retrieves processor internal and external data inputs, and processes them. Finally, the storage component stores the ingestion outputs in the cloud with the final format being Network Common Data Form (NetCDF) and all metadata stored in a PostgreSQL database. NetCDF was chosen because it's self-describing, machine-independent and well supported through software libraries for creation, access, and sharing of array-oriented scientific data.

It is then possible to gain remote sensing insights from both EO data and products, Fig. 3, in addition to the modelling forecasts and ground based data coming from the PET methodology and locally deployed weather/soil sensors (deployment in progress).

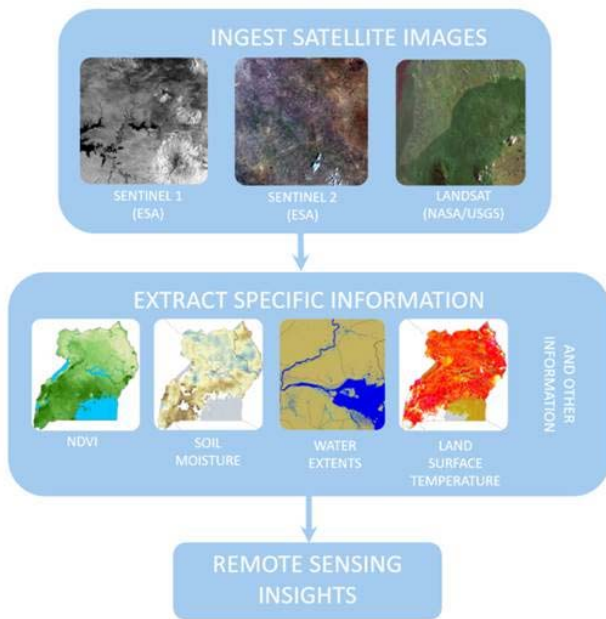


Fig. 3. EO data ingestion

The Visualization component allows DFMS users to access EMS services externally from the DFMS platform with a Graphical User Interface (GUI) web application (the geportal) and Public API.

2.4. Processing Approach for an Example Product: Soil Moisture

The soil moisture software architecture is broken down into the processing, data exchange, ingestion, and storage micro-services as shown in Fig. 4. The 'Produce soil moisture' step includes both the download of the original data and conversion to a daily composite. For version 1 (v1) it was just SMOS that was ingested while for v2 the input dataset is being expanded to potentially include ASCAT and SMAP alongside SMOS; depending on the pre-ingestion verification process. Version 1 was delivered in September 2017 and v2 is due for December 2018, with further iterations occurring throughout the length of the project up until December 2020.

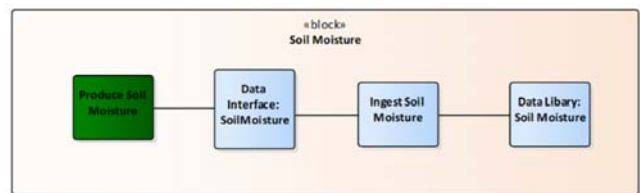


Fig. 4. Soil moisture software architecture

The pre-ingestion verification includes downloading all potential input soil moisture data sets and then comparing the EO data to modelled precipitation (primarily rainfall) from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim global atmospheric reanalysis [8] in v1, expanded to also include the Met Office East Africa (5°N to 15°S and east of 30°E) Model [9] in v2. The alignment is assessed in terms of the correlation between the datasets over the course of at least a year, with varying time lags being included. For the initial versions, the extracted points are just within the Karamoja region in Uganda's north-east corner that's the primary focus of the project, but soil moisture is being generated for the whole of Uganda and so the verification process will be expanded geographically.

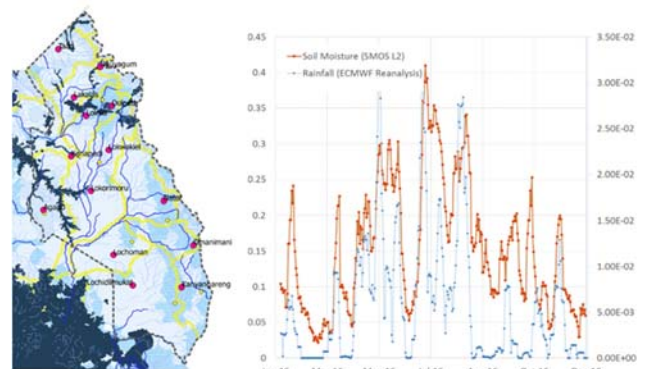


Fig. 5. Soil moisture pre-ingestion verification as the (left) location of the extraction points within the Karamoja region and (right) plot of soil moisture and rainfall.

An example v1 soil moisture product is shown in Fig. 6, with the NetCDF file containing both the daily average (shown) alongside the daily pixel count and daily standard deviation. The input Level 2 soil moisture data is currently the Near Real Time (NRT) Level 2 product generated using a Neural Network algorithm [10]; called SM-NRT-NN. These data are on the 15km spatial resolution Icosahedral Snyder Equal Area (ISEA) aperture 4 resolution 9 (4H9) global hexagonal grid; chosen for its characteristics of equal area and almost uniform intercell spacing [11]. The output Coordinate Reference System for DFMS is then Universal Transverse Mercator (UTM) zone 36 North at 1km resolution. Therefore, the non-square input pixels are transformed to square pixels, but while SMOS is the only input data the original pixel shape remains visible.

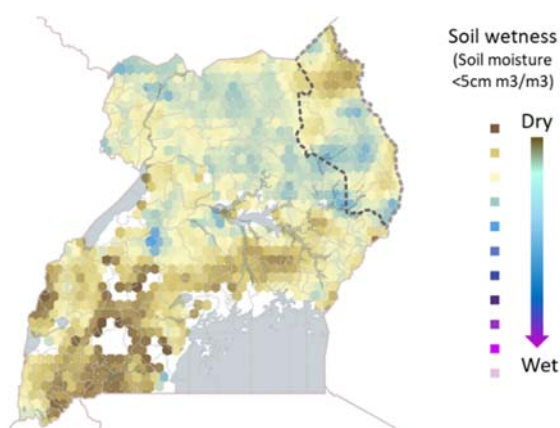


Fig. 6. Example soil moisture daily composite with the Karamoja region highlighted.

This pre-ingestion verification tool is also being expanded into a cross-verification tool that will automatically assess quality by looking for anomalies in dataset relationships.

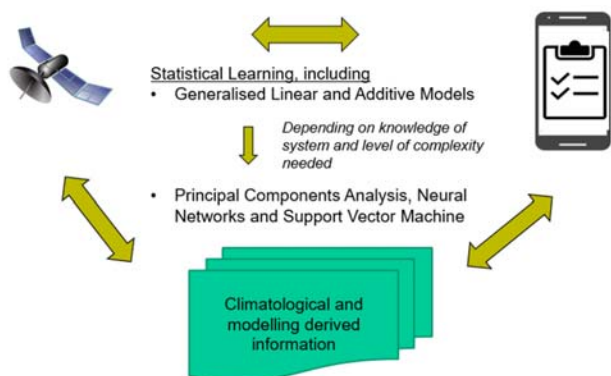


Fig. 7. Proposed approach for the cross-verification tool

3. DISCUSSION

The first of a number of visits by UK partners to Uganda took place in March 2017, where there was a workshop in

Kampala and one-to-one meetings that provided the opportunity to make local contacts and meet some of those whom we hope to benefit from this work. A further visit, to the district of Karamoja in the North of Uganda, occurred in July with a follow-up workshop in Kampala in September 2017 when an initial version of the platform and products was presented.

In parallel, the DFMS platform team is working towards a development release during the first quarter of 2018. The underpinning architecture is in place and EO datasets are being generated, and can be displayed within the ODC, alongside the localised weather forecasts. The next step is further integration of the input data streams so that platform starts to generated information that makes the most of all the available data streams.

4. CONCLUSIONS

Overall the DFMS platform aims to deliver output products with an improved quality, detail and frequency compared to those already available via other early warning platforms.

5. REFERENCES

- [1] <http://www.gou.go.ug/content/agriculture>
- [2] FAO, *The state of food and agriculture 2014: Innovation in family farming*, FAO, Rome, 2014.
- [3] <http://www.fao.org/uganda/fao-in-uganda/uganda-at-a-glance/en/>
- [4] D. Lumbroso. *Building the concept and plan for the Uganda National Early Warning System (NEWS)*. Final report. Evidence on Demand, UK (2016) xi + 82 pp. [DOI: http://dx.doi.org/10.12774/eod_cr.july2016.lumbrosod1], July 2016
- [5] <https://www.opendatacube.org/>
- [6] T. Cecere, "The Move Toward Analysis Ready Data and the Opportunities / Challenges Ahead," *JACIE 2016*, Fort Worth. https://calval.cr.usgs.gov/wordpress/wp-content/uploads/16.041-JACIE2016-Cecere_v3.pdf
- [7] <http://www.agritechtalk.org/PETmanualandmethodology.html>
- [8] D.P. Dee et al. "The ERA-Interim reanalysis: configuration and performance of the data assimilation system," *Q. J. R. Meteorol. Soc.*, 137: 553–597, 2011. DOI:10.1002/qj.828
- [9] [http://navigator.eumetsat.int/discovery/Start/DirectSearch/DetailedResult.do?f\(r0\)=EO:EUM:DAT:MODEL:4KM-EA](http://navigator.eumetsat.int/discovery/Start/DirectSearch/DetailedResult.do?f(r0)=EO:EUM:DAT:MODEL:4KM-EA)
- [10] N. J. Rodríguez-Fernández et al. "Soil moisture retrieval using neural networks: application to SMOS," *IEEE Transactions on Geoscience and Remote Sensing*, 53, 11, November 2015
- [11] M. Talone, Portabella, J. Martinez, and V. González-Gambau, "About the optimal grid for SMOS Level 1C and Level 2 products," *IEEE Geoscience and Rem. Sensing Lett.*, 12, 8, pp. 1630-1634, 2015.

6. ACKNOWLEDGEMENTS

The EC, ESA, NASA and the USGS for access to the input Earth observation data streams ingested into the DFMS platform alongside funding as part of the UK Space Agency's International Partnership Programme.

RHETICUS®: A CLOUD-BASED GEO-INFORMATION SERVICE FOR GROUND INSTABILITIES DETECTION AND MONITORING BASED ON FUSION OF EARTH OBSERVATION AND INSPIRE DATA

*Sergio Samarelli^a, Vincenzo Massimi^a, Luigi Agrimano^a, Daniela Drimaco^a,
Raffaele Nutricato^b, Davide Oscar Nitti^b, Maria Teresa Chiaradia^c*

^a Planetek Italia s.r.l., Via Massaua, 12, 70132 Bari, Italy;

^b Geophysical Applications Processing s.r.l., Via Amendola 173, 70126 Bari, Italy; Affiliation(s)

^c Department of Physics, Politecnico di Bari, Via Amendola 173, 70126 Bari, Italy.

ABSTRACT

The Rheticus® cloud-based platform provides continuous monitoring services of the Earth's surface. One of the services provided by Rheticus® is the *Displacement Geoinformation Service*, which offers monthly monitoring of millimetric displacements of the ground surface, landslide areas, the stability of infrastructures, and subsidence due to groundwater withdrawal/entry or from the excavation of mines and tunnels. To provide this information, the Rheticus® platform processes a large amount of Geospatial BigData. In particular, Rheticus® processes satellite Open Data provided by Copernicus Sentinels and it is capable to integrate local INSPIRE data sources. Rheticus® can automatically process the datasets that are compliant to the INSPIRE data specifications. We outline the automatic generation process of displacement maps and we provide examples of the detection and monitoring of geohazard and infrastructure instabilities.

Index Terms— Geo-analytics, data fusion, Geospatial Big Data, INSPIRE, Cloud computing, ESA Sentinel

1. INTRODUCTION

Geospatial information is today essential for organizations and professionals working in several industries. More and more, huge information is collected from multiple data sources and is freely available to anyone as open data. Rheticus® is the Planetek cloud-based data and services hub [1] able to process radar and optical data from multiple open-data satellite constellations designed to deliver updated geoanalytics information through complex automatic processes and minimum interaction with human beings. Among the peculiarities of Rheticus, there is the capability to integrate, in the processing algorithms, ancillary data, gathered in different ways from different data sources and channels, either for validation purposes or for increasing the information resolution enabled by the EO data used. “Rheticus displacement” represents a revolutionary model concept (through subscription) in monitoring Critical

Infrastructure (dams, pipelines, bridges etc) with the use of SAR data and Persistent/Distributed Scatterers technique (PS/DS), designed for users with high expectations in the value of information and its user friendliness provision.

Rheticus® Displacement is a powerful tool to prevent and detect potential sewerage failures (water networks, District heating), monitoring the wide-scope works like bridges, dams, buildings, even in relation to hydrogeological instabilities, landslides or the high traffic volume of metropolitan in the big cities.

The Rheticus® services through standardized and ready-to-use information is useful for a wide range of private and public organizations everywhere in the world, providing:

- Updated knowledge of ground and water conditions;
- Continuous monitoring of risk areas;
- Excellent Cost/Benefit relation;
- Safeguarding of investments.

To provide this information, the Rheticus® platform processes a large amount of Geospatial Big Data; it acts as an interoperable service node offering processing capabilities within a wider Big Data infrastructure. In particular, Rheticus® processes satellite Open Data provided by Copernicus Sentinels and it is capable to integrate local INSPIRE data sources like “building”, “administrative unit”, “hydrography”, “transport network” etc. Rheticus® can automatically process the datasets that are compliant to the INSPIRE data specifications. So the INSPIRE data specifications increase the throughput of the data processed, enabling an automatic process that improve the effectiveness or efficiency gains.

Actually, one of the engine of the Rheticus® workflow is the discovery datasets that use the INSPIRE discovery services to find the datasets relating to the survey AOI. Further improvements are expected when INSPIRE SDI will provide services in a streaming way, as the Copernicus Ground Segment already does, which flows data directly from the various Ground Segments to the users. In this way, the Rheticus® workflow can get automatically the datasets from the INSPIRE data streaming, avoiding users from the datasets loading.

2. RHETICUS® DISPLACEMENT SERVICE

More specifically, Rheticus® Displacement provides accurate information to monitor over time, through Multi-Temporal Synthetic Aperture Radar Interferometry (MTInSAR) techniques, movements occurring across landslide features or structural weaknesses that could affect buildings or infrastructures.

The availability of open data radar images acquired by the Sentinel-1 (S1) satellite mission (designed by the European Space Agency in the framework of the Copernicus programme) has fostered the implementation of a fully automatic MT-InSAR processing chain. Users gain reliable geo-information over wide areas in a blink of an eye, thanks to continuous satellite monitoring and overcoming difficulties and costs of field measurement campaigns.

In particular, Rheticus® browses and accesses (on a weekly basis) the products of the rolling archive of ESA S1 Scientific Data Hub; S1 data are then handled by a mature running processing chain, which is responsible of producing displacement maps immediately usable to measure with sub-centimetric precision movements of coherent Persistent Scatterers.

The selected MT-InSAR algorithm is SPINUA® (“*Stable Point Interferometry even in Un-urbanized Areas*”), developed by GAP srl, a POLIBA spinoff company, in collaboration with the Department of Physics of Bari and the CNR-ISSIA institute of Bari.

SPINUA® is a MTInSAR processing chain – whose flow chart is outlined in Figure 1 – designed to measure surface displacement, exploiting the phase difference between backscattered microwave signals of SAR images received from slightly different satellite orbits.

SPINUA® has been extensively tested and applied on both large urban areas and sparsely urbanized regions for monitoring both unstable areas (e.g. those affected by subsidence, landslides, post-seismic deformations) and unstable infrastructures (e.g. dams, bridges, roads, railways, pipelines, and so forth).

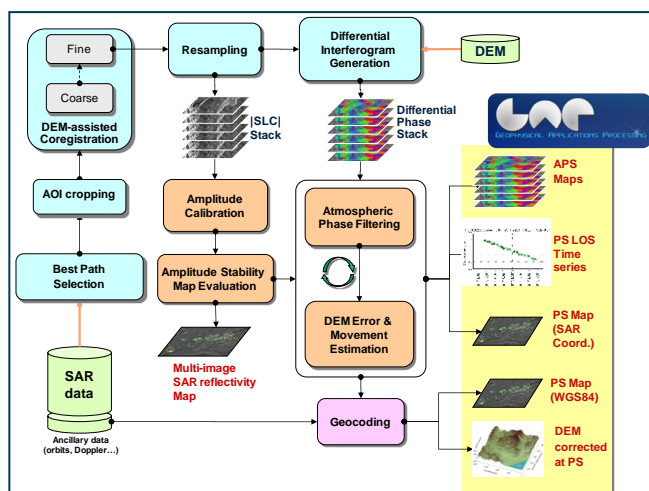


Figure 1. SPINUA® Flow Chart

Detailed information on the algorithmic solutions implemented in SPINUA can be found in [2,3]. In Figures 2÷6 we provide examples of the detection and monitoring of geohazards and infrastructure instabilities through the SPINUA algorithm.

Rheticus® has inherited the native and actual features of SPINUA, such as:

- integration of ground displacement maps obtained by processing radar images acquired by different sensors operating with different wavelengths and spatial resolutions;
- combination of displacements detected along ascending and descending passes for the estimation of the upward and eastward components of the ground movements;
- integration of Persistent and Distributed Scatterers;
- detection of non-linear movements, like accelerations and periodic seasonal trends;
- 3D building reconstruction from the PS spatial distribution.

Furthermore, Rheticus® Displacement service provides the user with graphic indicators, dynamic diagrams and preset reports (Figure 7) that allows assessing the level of stability of the monitored areas.

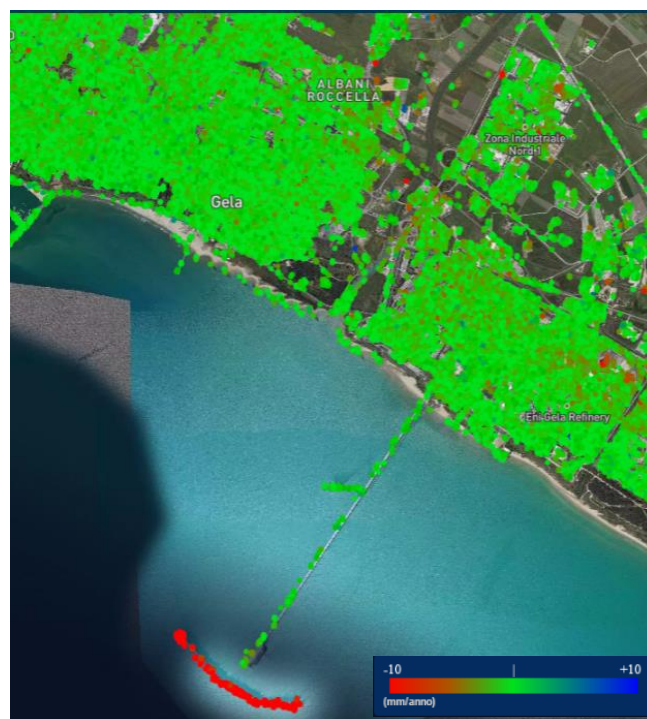


Figure 2. Offshore breakwater of the port of Gela (Sicily, Italy). The image shows the linear velocity measured along the radar Line-Of-Sight. Stable targets are marked by green dots. Red dots are representative, instead, of targets moving away from the satellite. Finally, blue dots show targets moving toward the satellite.

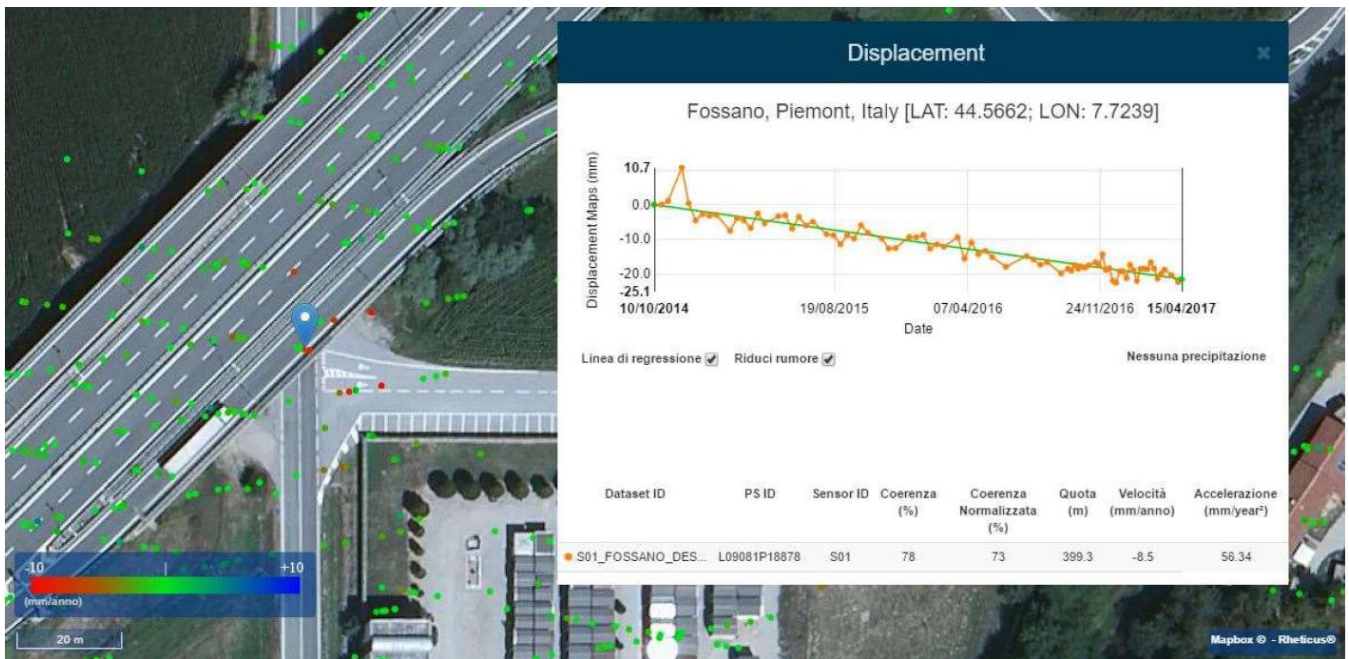


Figure 3. Infrastructure instabilities detected on the overpass close to the urban area of Fossano (Piemonte, Italy). Retrospective analysis on the detection of precursory signals related to the collapse occurred on April 2017.

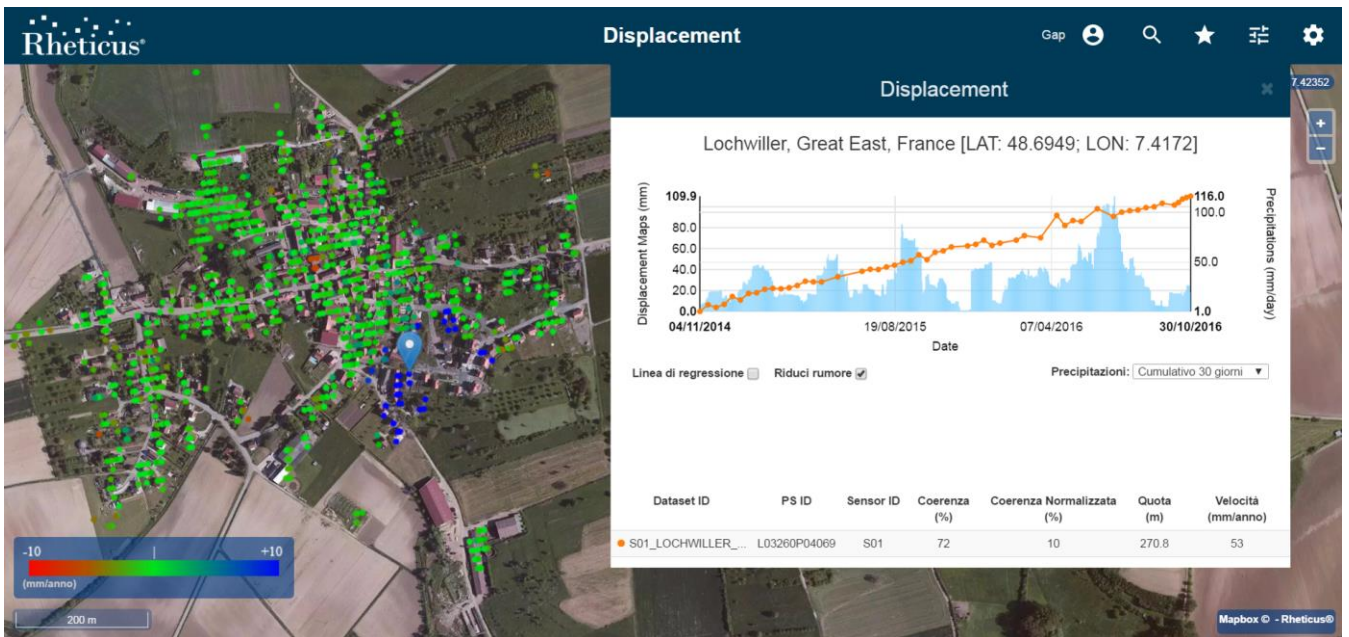


Figure 4. Uplifts monitoring in Lochwiller (France).

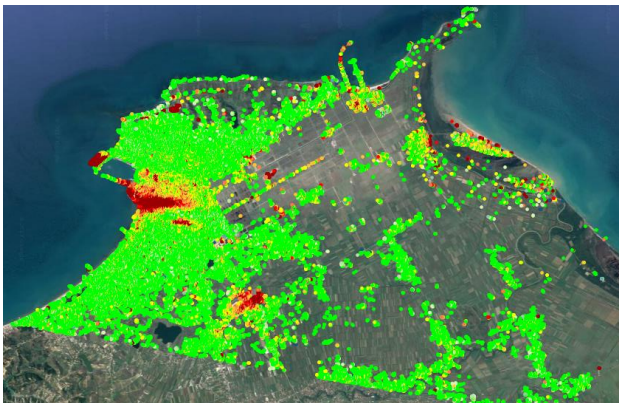


Figure 5. Subsidence monitoring over Durrës (Albania).

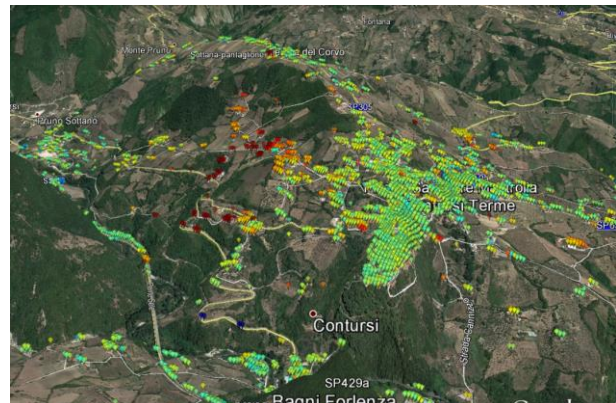


Figure 6. Landslide monitoring over Contursi (South Italy)

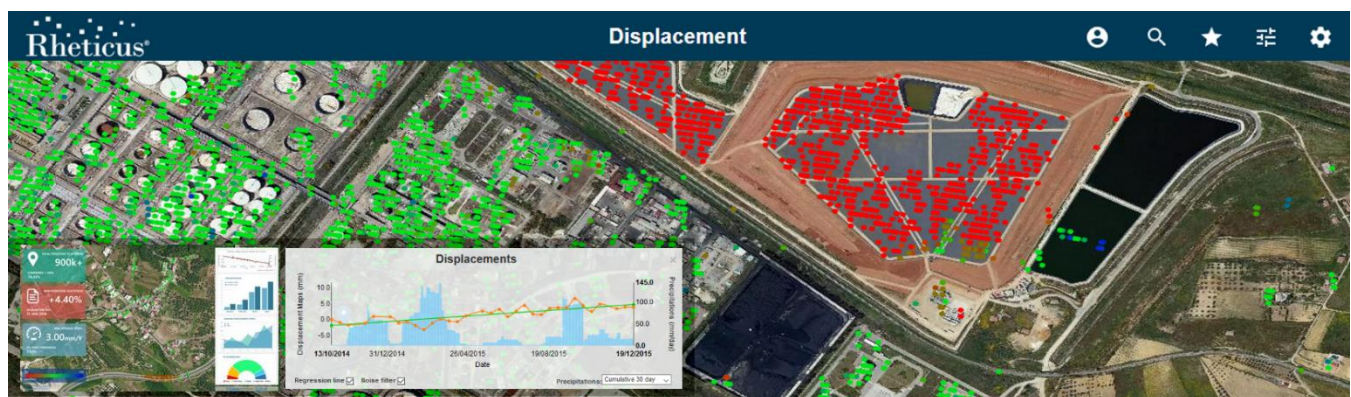


Figure 7. Rheticus® Displacement User Interface

3. FINAL COMMENTS

Rheticus® Displacement is an innovative, high-performing geo-information service for monitoring landslides, natural-and/or human-induced subsidence in urban and suburban areas, as well as stability of infrastructures (e.g. dams, bridges, buildings, road and railway networks, pipelines, electric towers, solar and wind farms, and so forth).

The service is accessible as cloud service and as web service following OGC standards. It is available in Machine-to-Machine mode (M2M) via standard sharing protocols.

It provides key parameters of surface displacement from satellite open data through extensively tested models and algorithms, and generates thematic maps, dynamic geo-analytics and pre-set reports.

It simplifies big data, streaming large information from various open data sources into an interactive and comprehensive dashboard to achieve insightful and purpose-built contents from many different perspectives.

Using European Copernicus Sentinel-1 (S1) open data images and MTInSAR techniques, Rheticus® Displacement service is complementary to traditional survey methods, providing a long-term solution to slope instability and geohazards monitoring.

4. ACKNOWLEDGMENTS

Rheticus® is a registered trademark of Planetek Italia srl. Research activities carried out in the framework of the FAST4MAP project (“Fast & Advanced SAR Techniques for Monitoring & Alerting Processes”) and co-funded by the Italian Space Agency (Contract n. 2015-020-R.0). Sentinel-1A products provided by ESA. Computational work partly executed on the IT resources made available by ReCaS grid and cloud infrastructure, a project financed by the MIUR (PON R&C 2007-2013).

5. REFERENCES

- [1] <http://www.rheticus.eu>
- [2] Bovenga, F., Refice, A., Nutricato, R., Guerriero, L. and Chiaradia, M. T., “SPINUA: A flexible processing chain for ERS/ENVISAT long term interferometry,” Proceedings of 2004 ESA-ENVISAT Symposium 1, 1-6 (2004).
- [3] Bovenga, F., Nutricato, R., Refice, A. and Wasowski, J., “Application of Multi-temporal Differential Interferometry to Slope Instability Detection in Urban/Peri-urban Areas,” Engineering Geology 88, 218-239 (2006).

SAR ALTIMETRY PROCESSING ON DEMAND SERVICE FOR CRYOSAT-2 AND SENTINEL-3 AT ESA G-POD

Jérôme Benveniste⁽¹⁾, Salvatore Dinardo⁽²⁾, Giovanni Sabatino⁽³⁾, Marco Restano⁽⁴⁾, Américo Ambrózio⁽⁵⁾,

⁽¹⁾ESA-ESRIN, Via Galileo Galilei, Frascati, Italy, Email: jerome.benveniste@esa.int

⁽²⁾He Space/EUMETSAT, ⁽³⁾Progressive Systems/ESRIN, ⁽⁴⁾SERCO/ESRIN,

⁽⁵⁾DEIMOS/ESRIN

ABSTRACT

The scope of this paper is to feature the G-POD SARvatore service to users for the exploitation of CryoSat-2 and Sentinel-3 data, which was designed and developed by the Altimetry Team at ESA-ESRIN EOP-SER. The G-POD service coined SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) is a web platform that allows any scientist to process on-line, on-demand and with user-selectable configuration CryoSat-2 SAR/SARin and Sentinel-3 SAR data, from L1a (FBR) data products up to SAR/SARin Level-2 geophysical data products. Several years of CryoSat-2 FBR data are at the disposal of the user, plus the full power of the G-POD's cluster: 600 CPUs and over 500 TB of storage.

Index Terms - SAR ALTIMETRY, CRYOSAT, SENTINEL-3, GPOD, SAMOSA, SENTINEL-3 STM

1. INTRODUCTION

The SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation (SARvatore) takes advantage of the G-POD (Grid Processing On Demand) distributed computing platform (600 CPUs in ~90 Working Nodes) to timely deliver output data products and to interface with ESA-ESRIN FBR data archive (380'000 SAR passes and 295'000 SARin passes for Cryosat-2). The output data products are generated in standard NetCDF format (using CF Convention), therefore being compatible with the Multi-Mission Radar Altimetry Toolbox (BRAT) and other NetCDF tools. By using the G-POD graphical interface, it is straightforward to select a geographical area of interest within the time-frame related to the Cryosat-2 SAR/SARin FBR and Sentinel-3 L1A data products availability in the service catalogue. The processor prototype is versatile, allowing users to customize and to adapt the processing according to their specific requirements by setting a list of configurable options. Pre-defined processing configurations (Ocean, Inland Water, Ice and Sea-Ice) are available for the Sentinel-3 service. After the task submission, users can follow, in real time, the status of the processing, which can be lengthy due to the required intense number-crunching inherent to SAR processing. From the web interface, users can choose to generate experimental SAR data products as stack data and RIP (Range Integrated Power) waveforms. The processing service, initially developed to support the awarded development contracts by confronting the deliverables to ESA's prototype, is now made available to the worldwide SAR Altimetry Community for research & development experiments, for on-site demonstrations/training in training courses and workshops, for cross-comparison to third party products (e.g. CLS/CNES CPP

or ESA SAR COP data products), for producing data and graphics for publications, etc. Initially, the processing was designed and uniquely optimized for open ocean studies. It was based on the SAMOSA model developed for the Sentinel-3 Ground Segment using CryoSat data (Cotton et al., 2008; Ray et al., 2014). However, since June 2015, a new retracker (SAMOSA+) is offered within the service as a dedicated retracker for coastal zone, inland water and sea-ice/ice-sheet. Following the launch of Sentinel-3, a new flavor of the service has been initiated, exclusively dedicated to the processing of Sentinel-3 mission data products. The scope of this new service is to maximize the exploitation of the Sentinel-3 Surface Topography Mission's data over all surfaces providing user with specific processing options not available in the default processing chain. The service is open, free of charge (supported by the ESA SEOM Programme Element) for worldwide scientific applications and available at https://gpod.eo.esa.int/services/CRYOSAT_SAR/. In this paper, we present first the ESA G-POD framework and system. Then we describe in detail the CryoSat-2/Sentinel-3 SAR Processing service integrated in G-POD and we conclude with the output package description and information on the contacts and references.

2. G-POD SYSTEM

The ESA Grid Processing on Demand (G-POD) system is a generic GRID-based operational computing environment where specific data-handling Earth-Observation services can be seamlessly plugged into system. One of the goals of G-POD is to provide users with a fast computational facility without the need to handle bulky data.

The G-POD system hosts high-speed connectivity, distributed processing resources and large volumes of data to provide scientific and industrial partners with a shared data processing platform fostering the development, validation and operations of new Earth Observation applications.

In particular, the G-POD environment consists of:

- Over 600 CPUs in about 90 Working Nodes
- Over 330 TB of local on-line Storage plus 180 TB of EO data accessed directly from the PACs.
- Access to Cloud processing and data resources on demand (from Interoute and other providers)
- Internal dedicated 1 Gbit LAN at ESA-ESRIN and at UK-PAC archives
- 1 Gbps external connection
- Online software resources: IDL, MATLAB, BEAT, BEAM, BRAT.

Actually, G-POD has more than 300TB of EO data locally stored. EO Data available to G-POD services come either from ESA and non-ESA missions. The G-POD web portal (<http://gpod.eo.esa.int/>) is a flexible, secure, generic and distributed web platform where the user can easily manage all own tasks. From the creation of a new task to the output publication, including data selection and job monitoring, the user goes through a friendly and intuitive user interface accessible from everywhere. More detailed information on the G-POD Web Portal and System are available here: <http://wiki.services.eoportal.org/tiki-index.php?page=GPOD+User+Manual#Annex>

3. CRYOSAT-2/ SENTINEL-3 SAR PROCESSING ON DEMAND SERVICE

The ESA G-POD Earth-Observation Service, SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) for CryoSat-2 and Sentinel-3 is an Earth-Observation application that provides the capability to process remotely and on demand CryoSat-2 SAR and Sentinel-3 data, from L1a (FBR, Full Bit Rate) data products until SAR Level-2 geophysical data products (Jensen and Raney, 1998; Wingham et al., 2006; Martin-Puig et al., 2008; Raney, 2008; Raney, 2012; Raney 2013).

The service works over any kind of surfaces but it has been so far optimized for ocean studies. It has been recently enhanced for inland water, land, sea-ice and ice sheets, implementing the SAMOSA+ model. The service is based on the SAR Processor Prototype that has been developed entirely by the ESA-ESRIN EOP-SER Altimetry Team (the authors) for CryoSat-2 & Sentinel-3 validation purposes, with the following system features:

- SAR/SARin FBR(L1a)/L1b DATA Archiving and Cataloguing
- SAR/SARin L1b Processor Prototype (Standard Delay-Doppler Processing)
- SAR/SARin L2 Retracker Prototype with SAMOSA Analytical Model and LEVMAR Least Square Estimator (Cotton et al., 2008; Ray et al., 2014)
- Input: CRYOSAT SAR/SARIN FBR DATA; Sentinel-3 SAR L1a Data
- Output L1b → Radar Echogram
- Output L2 → SSH, SLA (w/o SSB), SWH, sigma0, wind speed

The ESRIN EOP-SER ALT Team succeeded to compile the processor for a 64-bit Linux platform and delivered to the ESA G-POD team the executable codes, the input archive (CryoSat SAR FBR) and satellite footprints (ASCII tracks).

Now, the toolkit has been fully integrated in the GPOD System for gridded and on-demand computation.

The objectives of the service integration in GPOD are:

- to experiment in-house research themes that will be further matured in the ESA-funded R&D projects;
- to provide expert users with consolidated SAR geo-products to get acquainted with the novelties and specificities of SAR Altimetry;

- to validate CryoSat-2 & Sentinel-3 ocean products.

The service is open, free of charge and accessible online from everywhere. In order to be granted the access to the service, you need an EO-SSO (Earth Observation Single Sign-On) credentials (for EO-SSO registration, go to <https://earth.esa.int/web/guest/general-registration>) and afterwards, you need to submit an e-mail to G-POD team (write to eo-gpod@esa.int), requesting the activation of the service for your EO-SSO user account.

After the registration to EO-SSO, users can freely access the online service at: https://gpod.eo.esa.int/services/CRYOSAT_SAR/, https://gpod.eo.esa.int/services/SENTINEL3_SAR/. The services are listed under the Marine Theme. You can find them using the search bar as well.

The current GPOD service works only in SAR and SARin Mode (no LRM mode). As of October 2017, in the service catalogue, we have stored 380 thousands of SAR passes over the entire globe for period 2010-2017. This amounts to 160 TB of CryoSat-2 FBR data archived into G-POD storage. They can be all processed on-demand and online at user request.

4. WEB USER INTERFACE

Once you get to the service page (Fig. 1), the first action is to select the zone of interest and the time of interest for the required run. Regarding the selection of the area of interest, the user can simply draw a rectangle on the world map, after clicking on the rectangle icon on the tool bar. Instead, for more precise geo-selection, the user can type directly the geo-coordinates of the area of interest using the geographical bar.

Regarding the time of interest, the user may set the start and stop dates in the calendar bar. By default, the start date is the time of CryoSat-2/Sentinel-3 launch and the stop date is set at 2 months prior to the current date. The GUI embeds all the standard buttons for image browsing as panning, zoom-in zoom-out, centering, undo, redo, reset, etc.

Once the time and geo selection is done, clicking on the "QUERY" button, the service lists all the CryoSat-2/Sentinel-3 passes matching the time and space requirements. The CryoSat-2/Sentinel-3 SAR tracks, crossing the area of interest, are then shown on the world map in overlay. The graphical interface lists a maximum of 100 passes per page and informs users of the total number of found passes. The user can decide which passes to select by clicking on the passes, select all, or delete some specific passes from the list.

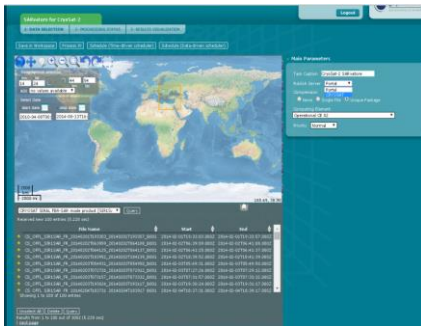


Figure 1: G-POD CryoSat-2 Service Main Interface.

On the top right, user finds a preference panel wherein user can set:

- Name of the current task
- Ftp Server where to publish the results (portal or personal)
- Data compression (tgz, none, single file)
- Grid Computing Resources
- Task Priority

The last step, before submitting the task, is to set the list of processing options. Indeed, the processor prototype is versatile in the sense that the users can customize and adapt the processing algorithms with flags and parameters, according their specific requirements, acting upon a list of configurable options. In the G-POD interface, users can easily enter this list of processing options via a series of drop-down menus. The configurable options are divided according to the processing level they refer to (L1b and L2). Starting from the first SARvatore release in 2014, the following upgrades have been introduced:

- Support for CryoSat-2 SARin Data.
- Enhancement of re-tracking capabilities in coastal zone and inland water by means of an advanced SAMOSA algorithm (SAMOSA+).
- Added support for posting rate at 80 Hz in delivering the output geophysical parameters.
- New Tide Model (TPX08) and Geoid (EGM2008, EIGEN-4C6).
- Support for Sentinel-3 SAR data.

Moreover, by selecting the processing options properly, users can mimic the CryoSat-2 or the Sentinel-3 processing baseline for an easy cross-comparison between missions. Pre-defined processing configurations (Ocean, Inland Water, Ice and Sea-Ice) are available for the Sentinel-3 service. Once the user has selected his processing options, in order to submit the task to G-POD Computing Elements, remains to click on the “PROCESS IT” button. After submission of a job, users will be directed to the workspace page where they can monitor in real time the status of the run and can be notified on the run status. The color code is:

- **Orange** → run under processing
- **Green** → run completed
- **Red** → run failed

Furthermore, by clicking on the task, the user can have more information on the processing task, such as:

- Task Id
- Task Creation Time
- Processing Id
- Grid Working Node Id
- Task Progress (retrieving, processing, publishing)

After run completion, by clicking on the button “Jobs Information”, the user can inspect:

- the GPOD log file (.stdout or .stderr) where eventual errors on data retrieving or data storing are reported;
- the prototype configuration file (L1b_CONFIG_FILE.log and L2_CONFIG_FILE.log) where are reported all the processing options;
- the prototype log files (L1b_start.log and L2_start.log) where are reported eventual prototype processing errors.

Users can also decide to change one or more processing options and then re-submit the task. In case of successful run completion (green status), the portal will provide an http link from where to download the output package on the user’s own local drive. The users can order to post the package directly on a personal ftp server after having communicated to the web platform the ftp server credentials (through the “publish servers” sub-menu). This is the recommended option in case of processing of large amount of data.

Future releases will:

- Support the UPorto GPD wet correction.
- Support the Tide Model (FES 2012).
- Provide a sea state bias solution.

5. OUTPUT PACKAGE & BRAT TOOLBOX COMPATIBILITY

The output package consists of:

- Satellite Pass Ground-Track in KML format
- Radar Echogram Picture in PNG format
- L2 Data Product in NetCDF format containing all the scientific results

The NetCDF format is self-explanatory with all the data field significance described in the attributes. The NetCDF Data Product follows the CF (Climate&Forecast) 1.6 Convention and can be opened with any standard NetCDF tools (ncdump, HDFview, etc.).

The following upgrades have been introduced for NetCDF Data Products:

- Inclusion of SAR echo and SAR RIP (Range Integrated Power) waveforms in the NetCDF files.
- Inclusion of STACK Data in the NetCDF files.

The recommended option is to ingest the NetCDF Data Products in BRAT Toolbox in order to exploit all the BRAT functionalities to browse and visualize the output content (Fig. 2). The Broadview Radar Altimetry Toolbox (BRAT) is a software suite designed to facilitate the use of radar altimetry data. It is able to read most distributed radar altimetry data, from ERS-1, ERS-2,

TOPEX/Poseidon, Geosat Follow-On, Jason-1, Jason-2, Envisat, CryoSat-2, Jason-3 and Sentinel-3, to perform some processing, data editing and statistics, and to visualize the results. As part of the Toolbox, a Radar Altimetry Tutorial provides information about radar altimetry, the technique involved and its applications, as well as an overview of past, present and future missions, including information on how to access data and additional software and documentation. It also presents a series of data use cases, covering all uses of altimetry over ocean, cryosphere, inland water and land, showing the basic methods for some of the most frequent manners of using altimetry data. BRAT has been developed under contract with ESA and CNES (<http://www.altimetry.info> and <http://earth.esa.int/brat/>).

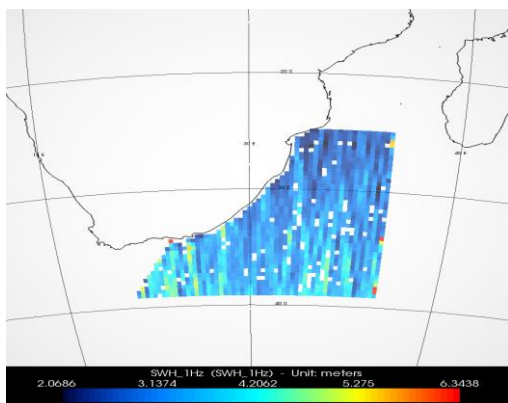


Figure 2: G-POD SAR CryoSat-2 Data products (Wave Height) opened in BRAT

6. CONCLUSIONS

To foster a new generation of SAR altimeter specialists and to get prepared for the Scientific Exploitation of Operational Missions (SEOM), a configurable versatile SAR processor has been developed and hosted in the ESA G-POD infrastructure. The G-POD Service coined SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) is a web platform that provides the capability to process on-line and on-demand CryoSat-2 and Sentinel-3 SAR data, from L1a (FBR) data products until SAR Level-2 geophysical data products, with a suite of selectable configuration parameters. The processing algorithms are the ones used in the Sentinel-3 Ground Segment, which mathematical model, SAMOSA, is described in Ray et al. (2014). By selecting the processing options properly, users can mimic the CryoSat-2 or the Sentinel-3 processing baseline for an easy cross-comparison between missions. Moreover, specific processing options not available in CryoSat-2 and the Sentinel-3 processing baselines have been made available to users. Pre-defined processing configurations (Ocean, Inland Water, Ice and Sea-Ice) are available for the Sentinel-3 service. The Broadview Radar Altimeter Toolbox can display the output of SARvatore. The service is open, free of charge and accessible online from everywhere.

7. FURTHER INFORMATION

For any question, bug report and support, please contact us at: altimetry.info@esa.int and eo-gpod@esa.int

Service Manual is available at:

<http://wiki.services.esa.int/tiki-index.php?page=GPOD+CryoSat-2+SARvatore+Software+Prototype+User+Manual>

SARvatore is available at:

https://gpod.esa.int/services/CRYOSAT_SAR/
https://gpod.esa.int/services/SENTINEL3_SAR/

BRAT is available at: <http://earth.esa.int/brat>

8. REFERENCES

- Cotton, D. et al., 2008, Development of SAR Altimetry Mode Studies over Ocean, Coastal Zones and Inland Water - State of the Art Assessment, <http://www.satoc.eu/projects/samosa/docs/SAMOSATN01-V1.0full.pdf>
- CryoSat User Workshop proceedings, SP-717, 2014, http://www.spacebooks-online.com/product_info.php?products_id=17581&osCsid=sscldrhw/
- Dinardo, S. and J. Benveniste, Guidelines for the SAR (Delay-Doppler) L1b Processing, ESA XCRY-GSEG-EOPS-TN-14-0042, Is. 2.3, 29/05/2013.
- Jensen, J. R., and R. K. Raney, "Delay Doppler radar altimeter: Better measurement precision," in Proceedings IEEE Geoscience and Remote Sensing Symposium IGARSS'98. Seattle, WA: IEEE, 1998, pp. 2011-2013.
- Martin-Puig, C. et al., SAR Altimetry Applications over Water, ESA SeaSAR Workshop, 21-25 January, SP-656, 2008.
- Raney, R. K., "CryoSat SAR-Mode Looks Revisited," Proceedings, ESA Living Planet Symposium, Bergen, Norway, 2010.
- Raney, R. K., "CryoSat SAR-Mode Looks Revisited," IEEE Geoscience and Remote Sensing Letters, vol. 9, pp. 393-397, 2012.
- Raney, R. K., "Maximizing the intrinsic precision of radar altimetric measurements," IEEE Geoscience and Remote Sensing Letters, vol. 10, pp. 1171-1174, 2013.
- Ray, C. et al., SAR Altimeter Backscattered Waveform Model, IEEE Trans. GeoSci. And Rem. Sens., Vol. 53, Iss. 2., pp 911 – 919, 2014. DOI: 10.1109/TGRS.2014.2330423.
- Wingham D. J. et al., CryoSat: A Mission to Determine the Fluctuations in Earth's Land and Marine Ice Fields. *Advances in Space Research* 37 (2006) 841-871.

9. UNIVERSAL RESOURCE LOCATORS (URL)

SEOM web site	http://seom.esa.int/
ESA Earth Online	http://eopi.esa.int/
Sentinels Online	http://sentinel.esa.int/
Copernicus	http://www.copernicus.eu
CP4O	http://www.satoc.eu/projects/CP4O/
Coastal Altimetry Workshops	http://www.coastalaltimetry.org
RADS	http://rads.tudelft.nl
SAMOSAT	http://www.satoc.eu/projects/samosa/
AVISO+	http://www.aviso.altimetry.fr/

BROADVIEW RADAR ALTIMETRY TOOLBOX

*A. Garcia-Mondéjar¹, R. Escolà¹, G. Moyano¹, M. Roca¹
M. Terra-Homem², A. Friaças², F. Martinho², E. Schrama³, M. Naeije³,
A. Ambrozio⁴, M. Restano⁵, J. Benveniste⁶*

1 isardSAT Ltd., 2 DEIMOS Engenharia, 3 TU Delft,
4 DEIMOS/ESRIN, 5 Serco/ESRIN, 6 ESA/ESRIN

ABSTRACT

The universal altimetry toolbox, BRAT (Broadview Radar Altimetry Toolbox), which can read all previous and current altimetry missions' data, incorporates now the capability to read the upcoming Sentinel-3 L1 and L2 products. ESA endeavoured to develop and supply this capability to support the users of the Sentinel-3 SAR Altimetry Mission.

Index Terms— BRAT, toolbox, altimetry, radar, Sentinel-3

1. INTRODUCTION

BRAT project started in 2005 from the joint efforts of ESA (European Space Agency) and CNES (Centre National d'Etudes Spatiales) and is a collection of tools and tutorial documents designed to facilitate the processing of radar altimetry data.

2. THE TOOLBOX

The toolbox enables users to interact with the most common altimetry data formats. Moreover, BRAT can be used in conjunction with MATLAB/IDL (via reading routines) or in C/C++/Fortran via a programming API, allowing the user to obtain desired data, bypassing the data-formatting hassle. However, BRAT can also be simply used to quickly visualise data or to translate the data in to other formats such as netCDF, ASCII text files, KML (Google Earth) and raster images (JPEG, PNG, etc.).

Several kinds of computations can be done with BRAT, involving combinations of data fields that the user can save for future uses or using the already embedded formulas that include the standard oceanographic altimetry formulas.

The BRAT Graphical User Interface (GUI) is the front-end for the powerful command line tools that are also part of the BRAT suite.

2.1 FUNCTIONALITIES

BRAT consists of several modules operating at different levels of abstraction. These modules can be Graphical User Interface (GUI) applications, command-line tools, and interfaces to existing applications (such as IDL and MATLAB) or application program interfaces (APIs) to programming languages such as C, FORTRAN and Python.

The main BRAT functions are:

- Data Import and Quick Look: basic tools for extracting data from standard formats and generating quick-look images.
- Data Export: output of data to the netCDF binary format, ASCII text files, or GeoTiff + GoogleEarth (KMZ/KML export); raster images (PNG, JPEG, BMP, TIFF, and PNM) of visualisations can be saved.
- Statistics: calculation of statistical parameters from data.
- Combinations: computation of formulas involving combinations of data fields (and saving of those formulas).
- Resampling: over and under-sampling of data; data binning.
- Data Editing: data selection using simple criteria, or a combination of criteria (that can also be saved).
- Exchanges: data editing and combinations can be exchanged between users.
- Data Visualisation: display of results (see Figure 1 and 2), with user-defined preferences. The viewer enables the user to display data stored in the internal format (netCDF).
- Download and periodic synchronization of satellite products with RADS database.

APIs are available with data reading, date and cycle/pass conversion and statistical computation functions for C, FORTRAN, IDL, (only using previous versions of BRAT), MATLAB and Python, allowing the integration of BRAT functionality in custom applications. For the most common

use cases (selection, combinations, visualisations, etc.), command-line tools are available that can be configured by creating parameter files. For beginners, we recommend using the BRAT GUI application, which enables the operator to easily specify the processing parameters required by each tool (and then invoke those tools at the push of a button).

BRAT is provided as Open Source Software, enabling the user community to participate in further development and quality improvement

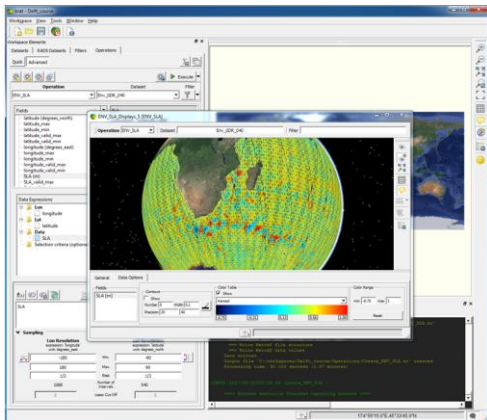


Figure 1 Envisat Sea Level Anomalies shown in BRAT.

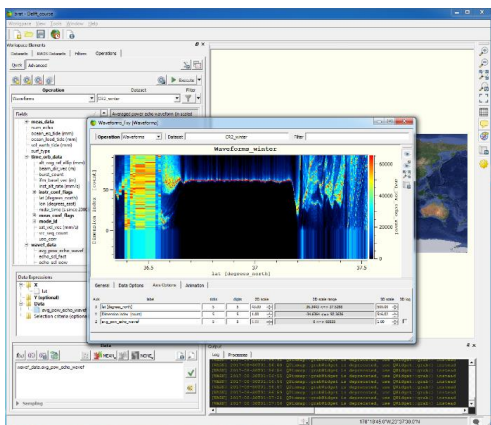


Figure 2 CryoSat-2 Waveforms over the Himalayas shown in BRAT

3. THE TUTORIALS

In the BRAT website, there is a set of different tutorials. The Radar Altimetry Tutorial, which is an update of the existing one at the beginning of the project, contains a strong introduction to altimetry and shows its applications in different fields such as Oceanography, Cryosphere, Geodesy and Hydrology, among others. On the other hand, the SAR altimetry tutorial has been created specifically for the current project in order to make the users aware of the great potential of SAR altimetry, specially coastal and inland applications.

Apart from these two tutorials, the user can find written and video tutorials showing how to use the toolbox and, at the same time, presenting some “use cases” for both conventional and SAR altimetry.

4. THE USER COMMUNITY

One of the main goals of the BRAT consortium is to create a user community around the project. Apart from periodically organizing webinars, workshops and trainings, the project has created a forum on the website (Figure 3 is a snapshot of the main page) where users can discuss on any matter and share their knowledge, and even share their modifications or additional parts of the toolbox’s code, which can be downloaded from the website as well.

Moreover, BRAT consortium has created a helpdesk, which is meant to be an interactive channel of communication with users of the toolbox or the tutorials. Some introductory and advanced video tutorials are available in our youtube channel

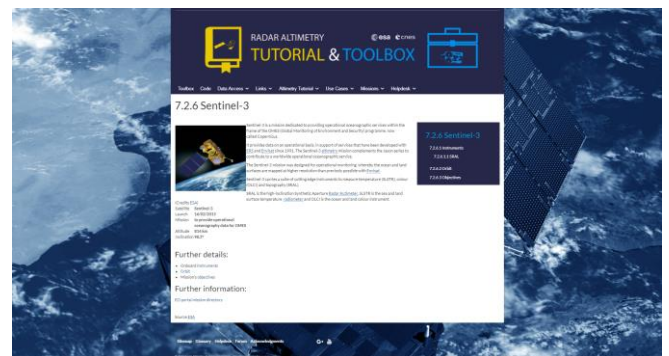


Figure 3 BRAT website with Sentinel-3 information

5. ACKNOWLEDGEMENTS

The Broadview Radar Altimetry Toolbox, under an ESA contract within the SEOM program, continues the work performed during 2006-2011 by CLS and S&T under and ESA and CNES contract.

If using the tutorial [1], please cite:

Rosmorduc, V., J. Benveniste, E. Bronner, S. Dinardo, O. Lauret, C. Maheu, M. Milagro, N. Picot, A. Ambrozio, R. Escolà, A. Garcia-Mondejar, M. Restano, E. Schrama, M. Terra-Homem, Radar Altimetry Tutorial; J. Benveniste and N. Picot (Editors).

6. REFERENCES

[1] The Radar Altimetry Tutorial, issue 1c, October 2016 - [link](#)

ATOS CODEX SPARKINDATA: PROMOTING USER UPTAKE OF AN EARTH OBSERVATION APPLICATION MARKETPLACE

J Emery¹, P.Lattes¹, C. Jaber², K. Konstantopedos²

¹Aerospace Valley, 118 Route de Narbonne, 31400 Toulouse, France

²Atos, 6 impasse Alice Guy, 31024 Toulouse, France

ABSTRACT

Bolstered by the arrival of free and open access to Copernicus data, space data derived applications are to bring substantial added value in the coming years. Valued at EUR 2.8 billion in 2015, the global Earth Observation (EO) downstream market is forecasted to increase to EUR 5.3 billion in 2020. European market share, estimated at EUR 632 million in 2015, is growing [1].

The European Commission highlighted in its 2016 Copernicus market report that the main innovation in the EO downstream market will be the implementation of platforms, drastically changing access to data.

As the European Commission (EC) moves towards the launch of its Copernicus Data and Information Access Services (DIAS), the emphasis in this paper is on the Atos Codex SparkInData (SID) project. Partially funded by the French General Commission for investment, SID is the first project of its kind to develop an EO exploitation platform.

SID, launched 3 years ago, today provides a case study in fostering user uptake of an EO application marketplace and boosting the creation of space data derived applications.

Index Terms— SparkInData (SID), Copernicus Data and Information Access Services (DIAS), Earth Observation (EO), Aerospace Valley (AV), Atos (ATOS), Centre National d'Études Spatiales (CNES) Institut National de l'Information Géographique et Forestier (IGN), Bureau des Recherches Géologiques et Minières (BRGM), Institut de Recherche Informatique de Toulouse (IRIT), European Space Agency Business Incubation Center (ESA BIC).

1. ATOS CODEX SPARKINDATA

The SID project is a partnership comprising: Atos as consortium leader, SMEs TerraNIS, Geomatys and Geosigweb, Mercator-Ocean, CNES, IGN, BRGM, IRIT at Paul-Sabatier University, the Engineering School of Purpan and the competitiveness cluster Aerospace Valley.

Its objective is to develop a user-oriented EO data access platform that overcomes technical and economic barriers preventing users from fully exploiting satellite data, enables the cross-fertilization of satellite data with other data sources and fosters the emergence of new services in downstream markets.

Together SID partners worked to build business-oriented, high-level services in various fields, ranging from Agriculture and Territorial Collectivities to Energy and Oceanography through complementary partners TerraNis, Geomatys, Geosigweb and Mercator-Ocean.

SID is a Marketplace:

- End users discover added-value Earth Observation data and services helping them to improve their day-to-day business.
- Data and Service providers monetize their Earth Observation data and services.

SparkInData is also an ecosystem allowing EO Data and Service providers to combine data and algorithms from tiers in order to create new added-value data for end-users.

SID solution (Fig. 1) is based on a:

- IaaS able to provide sufficient Computing and Storage resources following the cloud model (Scalability, Elasticity, on demand model)
- PaaS that
 - Interconnects, homogenizes and facilitates access to data and metadata.
 - Provides common base services for the higher level services to work.
 - Provides EO core components needed by all business oriented services to work.
 - Allows flexible and easy creation for the added value, business oriented and end-user services.
- SaaS allowing end-users to discover and use all the available EO data and services through the Marketplace where they are all exposed:

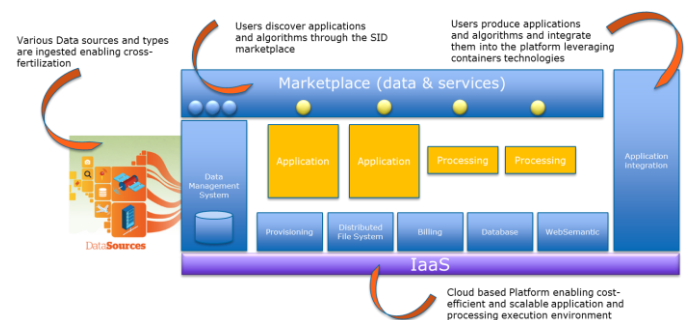


Fig.1 SID's functional architecture

1.1. The IaaS layer

Although the SID software components are totally cloud agnostic, IT resources deployed by the platform are virtual machines (VM) that are today spawned on the Atos Canopy Orchestrated Hybrid Cloud [2]. SID aims to be fully independent from the cloud provider thus it is possible to use any cloud provider resources or even to deploy it on a bare-metal machine cluster.

1.2. The PaaS layer

In the development of SID PaaS we have identified the following challenges:

- Failover Handling
- Scaling (including auto-scaling)
- Fast deployment
- Resources sharing
- Multitenancy (Authentication and Isolation)
- Big data handling
- Accurate billing of resources used

Furthermore SID shall be able to easily integrate and host all our partners' EO services. Each of these partners will potentially develop services in different languages or using different frameworks.

The PaaS architecture of SID has been built from the ground up in order to answer these challenges.

1.2.1. Service based Architecture

In order to cope with all the challenges listed previously, SID was built using a container based microservices and multitenant architecture. In fact, all features and functionalities are encapsulated in Services with the least possible coupling between them.

Since 2013, containers technologies have gained momentum enabling virtualization features offered by the Linux kernel. These features have been around for a long time now but projects like Docker have created an ecosystem where containers are easier to create and use [3]. A container provides an abstraction layer between the hosting Operating system and the application running inside the container. Unlike the traditional fully-fledged virtual machines (VM), containers need only to embed the required libraries and applications to run and leverage the hosting machine OS kernel to function normally. This additional abstraction reduces the coupling between the application and the machine on which it is executed and reduces drastically deployment time (less than a minute in average).

Building on this application container approach, the scalability, resilience, isolation and deployment are handled by container orchestration modules. This type of modules provides both enough isolation between the deployed

containers and a secure execution environment. Many orchestration frameworks have been introduced with different philosophies and each one of them has its pros and cons. SID's choice landed on the Google backed technology: Kubernetes [4]. This latter introduces concepts allowing in particular:

- Simple deployment and organization of containerized applications across multiple hosting machines.
- Applications interactions are also made easier by providing discovery mechanisms as Domain Name System (DNS).
- Storage Management (permission control, quotas, ...).
- Containers scheduling enables smart and **fine grained resources use and sharing** between applications.
- Application failover and rescheduling in case of machines or application failure or termination.

1.2.2. Common Services

SID features have been designed and thought to simplify EO data mining and handling. Common services (e.g. the services management service, data management service, authentication /authorization service, resources management service, computing framework service, billing service,...) with multitenancy support are provided off-the-shelf.

All these common services are accessible from the partners' EO services through specific REST APIs. This has the advantage to make service development much quicker and focused on the real added-value.

1.2.3. Service integration

As explained above, the platform has been conceived following a service-oriented architecture. Therefore, all partners and users intending to participate and make accessible their applications to the rest of the platform shall also deliver their applications following this concept of Service. Following this principle, applications can interconnect and benefit from each other creating the **SID ecosystem**.

In order to integrate new services or calculation algorithms, SID adopts two main approaches:

- The Service Manager
- The Processing Pipeline

The **Service Manager** allows defining and deploying full applications on SID. The developer defines:

- Each component (embedded in Dockers) of its application
- Physical resources (e.g. CPU, RAM, storage) needed.
- Auto-scaling triggers for one or several of his/her application components.

For its part, the service manager enforces the constraints on interface (e.g. documentation), on physical resources and makes all services discoverable. The deployed applications

can be published on the marketplace in order to be used from final users.

The **Processing Pipeline** approach rests on the WPS OGC standard [5], particularly on the open source GeoServer product [6] which is a compliant implementation of several OGC standards. The Processing Pipeline can be seen as a software layer on top of GeoServer automatizing the integration of new processing blocks and the generation of a WPS compliant documentation/interface. Through a simple GUI or a REST API, SID's users can define processing blocks from a given set of inputs parameters, outputs and the Docker image containing the algorithm (previously pushed to the platform registry). Next, the service takes care of generating the WPS interface needed to discover, execute and monitor the newly added processing. The choice of the WPS standard opens also the door to chain processing blocks.

1.2.4. PaaS Monitoring

The PaaS monitoring system is essential in a cloud based platform. The metrics and events occurring on the platform are collected not only to watch the platform state but also to automatize failover and scalability. These latter aspects are essential in order to implement a cost-efficient platform capable of handling the storage and compute loads expected of a platform such as SID. The monitoring system is implemented in particular with Influx Data stack, an open source time series platform built specifically for metrics, events, and other time-based data [7].

1.3. The SaaS layer

The end-user accesses and instantiates the different published services through a user-friendly interface called: **the marketplace**. The user is then redirected to the service interface (if it exists) or to common workspace interface allowing to call any provided service API and to visualize conveniently the results. In addition to data and PaaS level software blocks, SID partners works on the building of business oriented high level services in various fields going from Agriculture and territorial collectivities to Energy and Oceanography.

2. TERRANIS

TerraNIS provides an interesting success story for the SID project. Expert in imagery processing for agricultural applications, TerraNIS has been able to directly benefit from SID infrastructure, data, algorithms and visibility to develop its services. Capabilities such as data analysis and image processing directly via the platform have enhanced the SME's ability to process a large quantity of data. Moving forward, TerraNIS looks to develop an industrial solution based on SID capabilities beyond the scope of the project.

3. BOOSTER NOVA

To promote user uptake beyond the scope of the SID consortium, the 'market pull' approach undertaken by Aerospace Valley, via the Booster NOVA alliance, also provides an interesting case study.

Aerospace Valley is a competitive cluster bringing together more than 800 entities of the Aerospace sector, including 500 PMEs. It has a strong background in the promotion of Earth Observation technologies and data towards its regional ecosystem. Home to major space industry players and a plethora of innovative SMEs and start-ups, the region has strong potential regarding the development of applications and services.

Since early 2016, Aerospace Valley manages one of the 4 national 'Booster' initiatives launched by the French government: a leading cross-sectorial alliance named Booster NOVA that covers the urban areas of Montpellier, Bordeaux and Toulouse. Each partner of this cross sectorial coalition represents a specific value chain: Space data, Internet of Things, Digital and Big Data, but also agriculture, energy, blue growth, mobility, smart cities, developing countries... SID plays a front role as the big data partner within the alliance, key for detecting SMEs and/or projects for SID.

The objective of the alliance is clear: to foster new business services using a combination of space based EO data with other sources of data (open data, IOT, commercial data) for businesses. To fulfil such an ambition, Booster NOVA coordinates a seamless process called Dream it –Make it –Boost It to co-design, create and accelerate the development of these space-based services. Several success stories are highlighted to illustrate this philosophy below.

3.1. Dream-it: Fostering ideas, synergies, knowledge and know-how

3.1.1. Booster Nova Innovation Clubs

To leverage innovative services, workshops between stakeholders from application areas (agriculture, energy ...), technology providers and data specialists are organized. These innovation club use a new, innovative methodology drawing on research activates led by Aerospace Valley together with ESTIA University.

3.1.2. Booster Nova Meetup Events

To foster synergies, ideas, knowledge and know-how, regular open, informal events are organized by Aerospace Valley in partnership with La Mêlée, a digital business network based in South West France. Space4Digital events provide training and inform the wider digital community about the potential of space data, whilst Digital4Space events focus on the perspectives and opportunities of digital technology as a means to transform space data into viable applications.

3.1.3. ActInSpace

Imagined by the CNES and co-organized by ESA since its 2nd edition, this international space applications hackathon designed for students but open to everyone aims to stimulate creativity and student entrepreneurship [8]. 4 start-ups were created in 2014, 14 startups in 2016. 50 hosting cities, 24 countries and at least 2500 applicants are expected for its third edition in 2018.

3.1.4. FabSpace 2.0

This H2020 project looks to create a network for geodata-driven innovation, leveraging space data in Universities 2.0. It aims at launching GIS open innovation one-stop shops (FabSpaces) in European universities and European Space Agency Business Incubation Centers (ESA BICs). A platform has been developed as an ideation tool aiming to foster startup creation [9].

3.2. Make-it: Providing business support, expertise and funding for project maturation

3.2.1. PIAVE (*Industrial projects building the future*) *dedicated to space services*

This French government grant fund is dedicated to financing services using space data. Companies can finance projects during 2 different project stages: project maturation and industrialization. Aerospace Valley, via Booster Nova, has been mandated to support project proposals and fast track project applications.

3.2.2. 'Digital Challenges' by Booster Nova

These challenges are open innovation contests where customers provide technical or business needs to be solved by startups and SMEs. 8 Challenges have been defined in 3 key application sectors: Smart City, Agriculture and developing countries. Laureats from these challenges have privileged access to data but also to a 70k euros grant per challenge [10].

3.2.3. Neptune

This H2020 project looks to support the creation of new services in the blue growth field and develop new cross-sectoral industrial value-chains. 75% of the total budget is allocated to financing support actions for SMEs [11].

3.2.4. ESA BIC Sud-France

Managed by Aerospace Valley, this dedicated European Space Agency Business Incubation Center provides support to entrepreneurs and startups managers looking to deepen their use of space data and technology. Access to grants and to hours of technical expertise are available. The service offering is currently being reinforced to include EO. It has provided support to 48 incubated businesses to date [12].

3.3. Boost-it: Accelerating the commercial deployment of services and products

Aerospace Valley coordinates Space4Globe, the unique alliance of EU Clusters which works to identify international markets specifically for EO applications and accelerate their commercial deployment into these markets [13].

4. CONCLUSION

The SID platform has been a pioneer in facilitating access to (EO) data and information, boosting user uptake, stimulating innovation and exploring new business models. This experience can be shared and drawn upon as the Copernicus Data and Information Access Services (DIAS) platforms get underway to foster European and international user uptake.

5. REFERENCES

- [1] European Commission, PWC, "Copernicus Market report" Issue 1, November 2016
- [2] <https://atos.net/en/solutions/atos-canopy-orchestrated-hybrid-cloud/>
- [3] <https://www.docker.com/>
- [4] <https://kubernetes.io/>
- [5] <http://www.opengeospatial.org/standards/wps>
- [6] <http://geoserver.org/>
- [7] <https://www.influxdata.com/>
- [8] <http://www.actinspace.org/>
- [9] <https://www.fabspace.eu/>
- [10] <http://booster-nova.com/les-challenges-numeriques/>
- [11] <http://www.neptune-project.eu/>
- [12] <https://spacesolutions.esa.int/business-incubation/esa-bic-sud-france?language=en>
- [13] <http://www.space2id.eu/>

Author Index

Abbattista, Cristoforo	25, 363
Adam, Fathalrahman	293
Adeline, Karine	437
Agrimano, Luigi	473
Aimé, Stephan	328
Albani, Mirko	154, 211, 214, 217
Albani, Sergio	275
Aldeborgh, Nikki	255
Alfonsi, Lucilla	363
Alhaddad, Bahaaeddin	259
Ali, Iftikhar	9, 40
Ambrózio, Américo	477, 481
Amler, Esther	461
Amodio, Angelo	449
Amoruso, Leonardo	25, 363
Angiuli, Emanuele	275
Antonucci, Ester	359
Arbab-Zavar, Banafshe	445
Arcorace, Mauro	102
Arino, Olivier	157
Arviset, Christophe	193, 229
Asamer, Hubert	243
Aspetsberger, Michael	344
Assis, Luiz Fernando	48
Auburn, John	469
Augustin, Hannah	173
Bach, Heike	461
Bacsárdi, László	410
Baillarin, Simon	279
Baines, Deborah	229
Bairat, Musab	390
Balasis, Georgios	421
Balhar, Jakub	243, 414
Baraldi, Andrea	17
Barbarisi, Isa	229
Baroux, Christophe	332
Barrau Huguet, Sylvie	344
Barreyre, Clémentine	118
Basset, Antoine	383
Bauer-Marschallinger, Bernhard	40
Baumann, Peter	32
Baynes, Kathleen	197
Bednarczyk, Michał	336
Beebe, Reta	185
Belabbess, Badre	390
Belhadj-Aissa, Aichouche	398
Bellec, Nicolas	371
Bemporad, Alessandro	359
Bencsik, Gergely	410
Benveniste, Jérôme	477, 481
Beqiraj, Gudar	429
Bereta, Konstantina	94

Berjaoui, Ahmad	348
Berry, David	325
Bertoluzza, Manuel	114
Besombes, Jérôme	437
Besse, Sebastien	229
Bettge, Anika	301
Biasutti, Roberto	211
Blanc-Paques, Pierre	356, 387
Boettcher, Martin	243
Boissier, Enguerran	243
Bonano, Manuela	209
Bouchemakh, Lynda	398
Boulch, Alexandre	305, 352, 437
Bourrienne, Ségolène	387
Boussouf, Loïc	118
Bovolo, Francesca	114
Boyd, Ian	457
Braakmann-Folgmann, Anne	297
Brahem, Mariem	189
Briese, Christian	9
Brooks, Andrew	154
Brown, Lee	29
Brunet, Pierre-Marie	321
Bruzzo, Lorenzo	13, 114
Burger, Armin	52, 71, 271
Bushati, Salvatore	429
Cabon, Bertrand	118
Cado Van der Lelij, Amrit	247
Caillet, Claire	225
Canty, Morton J.	126
Cao, Senmao	9, 40
Carbone, Marianna	363
Cartus, Oliver	157
Carty, Shane	457
Casadio, Stefano	56
Castaigns, Thibaut	305
Castel, Fabien	21, 445, 465
Castelli, Elisa	56
Casti, Marta	359
Castracane, Paolo	457
Casu, Francesco	209
Catalao, Joao	375
Catarino, Nuno	453
Cavadini, Salvador	146
Cavallaro, Anna-Maria	173
Ceamenos, Xavier	437
Ceriola, Giulio	344
Cesaroni, Claudio	363
Chambers, Caroline	469
Chan-Hon-Tong, Adrien	352, 437
Chang, George	417
Chaoul, Laurence	106
Chapel, Laetitia	371
Chapron, Bertrand	169
Chaumat, Laure	1
Chausserie-Laprée, Benoît	225

Chen, Jiquan	449
Cherrier, Noëlie	305
Chiaradia, Maria Teresa	473
Chini, Marco	67
Churchill, Jonathan	287
Ciancarelli, Carlo	394
Cianchini, Gianfranco	363
Cigna, Francesca	165, 425
Clarke, Sebastian	469
Clerc, Sébastien	344
Colangeli, Guido	98
Colapicchioni, Andrea	461
Cooper, Lauren	449
Corbane, Christina	52, 267, 367
Correndo, Gianluca	445, 465
Cosac, Razvan	217
Costa, Armin	205
Craglia, Max	32
Crichton, Daniel	185
Cronin, Abigail	453
Cuccu, Roberto	340
Cuomo, Antonio	461
Curé, Olivier	390
Câmara, Gilberto	48, 441
d'Andrimont, Raphaël	367
Daems, Dirk	150
Daglis, Athanassios	421
Dahlin, Kyla	449
Daniel, Sandrine	29
Danne, Olaf	59
Datcu, Mihai	289, 293
De Franceschi, Giorgiana	363
De Groof, Arnaud	239
De Luca, Claudio	209
De March, Ruben	309, 379
De Marchi, Davide	71, 271
De Marchi, Guido	229
De Santis, Angelo	363
De Santis, Anna	363
de Teodoro, Pilar	193
de Vries, Mindert	247
Dean, Andy	461
Decoster, Nicolas	75
Deffacis, Maurizio	379
Defourny, Pierre	157
Dehn, Angelika	457
Delaney, Conor	457
Delgado Blasco, José Manuel	67, 102, 340
Demir, Begüm	13, 406
Dhu, Trevor	32, 36
Di Giovambattista, Rita	363
Dijkstra, Jasper	247
Dijkstra, Lewis	52
Dillmann, Martin	138
Dinardo, Salvatore	477
Dinelli, Bianca Maria	56

Diprima, Francesco	25
Dislaire, Jean-Christophe	332
Donadieu, Joëlle	279
Dorgan, Sébastien	328
Doyle, Chris	469
Dries, Jeroen	150
Drimaco, Daniela	363, 473
Dufour, Guillaume	437
Duprat, Stephane	134
Eggleston, James	283
Ekkelenkamp, Rudie	247
El Maalem, Driss	134
Elefante, Stefano	9, 40
Elfassi, Adrien	348
Emery, Joanna	483
Engdahl, Marcus	157
Esch, Thomas	243
Escolà, Roger	481
Espe, Thomas	457
Eynard-Bontemps, Guillaume	5
Fablet, Ronan	169, 371
Farquhar, Clive	239
Faudi, Jean-François	387
Favot, Laurent	279
Ferreira, Karine	48
Ferrer, Marc	134
Filippi, Fabio	359, 379
Fineschi, Silvano	359
Fjaeraa, Ann Mari	457
Florczyk J., Aneta	52
Folco, Sergio	154
Fonti, Andrea	359, 379
Fortunato, Vito	25
Fraisse, Renaud	356
Fraseri, Neki	429
Friaças, Ana	481
Friguet, Chloé	371
Fujita, Naoyuki	317
Gabet, Laurent	146, 387
Gale, Leslie	142
Garcia, Vincent	134, 344
Garcia-Mondejar, Albert	481
Garello, René	371
Gaucher, Julien	75
Gaudissart, Vincent	328
Gavin, David	36
Geoffret, Xavier	321
Georganos, Stefanos	263
Giamini, Sigiava	421
Giannakis, Omiros	421
Gibson, George	469
Gil, Amaia	375
Gilles, Nicolas	344
Gilliams, Sven	461

Ginet, Patrick	321
Gloaguen, Pierre	371
Gobron, Nadine	59
Goncalves, Romulo	63
Goor, Erwin	150, 461
Greco, Fedele	56
Grippa, Tais	263
Grosso, Nuno	453
Gutfreund, Denis	321
Hajduch, Guillaume	371
Hamilton, Steve	449
Hansen, Johannes	457
Hanson, Mary	402
Hardman, Sean	185
Harrison, Mark	469
Harwood, Phillip	461
Hasenohr, Paul	271
Healy, Paul	469
Heinen, Torsten	313
Held, Alex	32
Hendriksen, Gerrit	247
Hirner, Andreas	243
Hirtl, Marcus	236
Hoersch, Bianca	211
Hogan, Patrick	79
Horrein, Pierre-Henri	169
Hostache, Renaud	67
Hudson, David	36
Huesler, Fabia	154
Hughes, John S	185
Häme, Tuomas	239
Hämäläinen, Jarno	239
Iapaolo, Michele	142
Ikehata, Yousuke	317
Innocente, Luca	469
Intelisano, Arturo	394
Ippolito, Alessandro	363
Ivall, Thomas	445
Izquierdo Verdiguier, Emma	63
Jaber, Chadi	483
Jacob, Alexander	205
Janez, Fabrice	437
Jeanjean, Hervé	211
Jenkins, Elma	469
John, Ranjeet	449
Jones, Richard	469
Julie, Bastien	279
Karantzalos, Konstantinos	181
Karmas, Athanasios	181
Karnieli, Arnon	449
Kellndorfer, Josef	157
Kempeneers, Pieter	71, 177, 271
Kemper, Thomas	52

Kepeklian, Gabriel	390
Kershaw, Philip	239, 287
Kershaw, Trent	36
Kesa, Maria	259
Kettig, Peter	383
Keßler, Torben	201
Kharbouche, Said	59
Khrouf, Houda	390
Kiemle, Stephan	201
Kiernan, Paul	457
Killough, Brian	32
Kim, Richard	417
Klein, Julian	29
Koeniguer, Elise	437
Kolotii, Andrii	122
Konstantopedos, Kyriakos	483
Koubarakis, Manolis	94
Koukouvinos, Stathis	21
Kourgli, Assia	398
Kristen, Harald	205
Kroupi, Eleni	259
Kunze, Markus	201
Kuo, Kwo-Sen	90
Kusche, Jürgen	297
Kussul, Nataliia	122
L'Helguen, Céline	279
Lafortezza, Raffaele	449
Lally, Regina	469
Lamy-Thepaut, Sylvie	44
Lanari, Riccardo	209
Landgraf, Guenther	79
Lang, Stefan	17, 173
Lattanzi, Mario	309
Lattes, Philippe	483
Laur, Henri	211
Laurent, Béatrice	118
Lavender, Andrew	469
Lavender, Samantha	469
Lavreniuk, Mykola	122
Law, Emily	185, 417
Lawrence, Bryan	287
Lazecky, Milan	161
Lazzarini, Michele	275
Le Carvennec, Arnaud	469
Le, Tuan Sy	9
Lee, Young	29
Lefèvre, Sébastien	251, 371
Lemoine, Guido	367
Lennert, Moritz	263
Leone, Rosemarie	154, 217
Leroy, Marc	279
Leuzzi, Chiara	379
Lewis, Adam	32, 36
Lguensat, Redouane	169
Lhez, Jérémy	390
Lisima, Yanis	332

Llapa, Eduardo	48
Loekken, Sveinung	233
Lonie, Neil	154
Lopes, Cristiano	236
Lorenzo, Jose	445
Loubes, Jean-Michel	118
Louge, Camille	134
Lowe, Dawn	221
Lozano, Javier	433
Lumbroso, Darren	469
Lusch, Dave	449
Lymburner, Leo	36
Lynnes, Christopher	197
Maffenini, Luca	52
Maggio, Iolanda	217
Malhotra, Shan	417
Manganiello, Claudio	102
Mantovani, Simone	236
Manunta, Michele	209
Manzo, Mariarosaria	209
Marchetti, Dedalo	363
Marchetti, Pier Giorgio	102
Marconcini, Mattia	243
Marin, Alessandro	233
Marmanis, Dimitrios	289, 293
Marques, Gustavo	283
Marrazzo, Massinmo	102
Martinez, Beatriz	229
Martinho, Fernando	481
Martino, Michele	359
Masse, Antoine	225
Massey, Neil	287
Massimi, Vincenzo	473
Masson, Arnaud	229
Mateus, Pedro	375
Matgen, Patrick	67
Mathot, Emmanuel	243
McInerney, Mark	197
McKinstry, Alastair	457
Melcot, Matthieu	142
Melet, Olivier	5
Mercirol, François	251, 371
Merin, Bruno	229
Messineo, Rosario	309, 359, 379
Metz, Annekatrin	243
Michel, Aurélie	437
Migdall, Silke	461
Miguens, Joana	138
Mikai, Hidekazu	317
Milcinski, Grega	79
Milillo, Giovanni	425
Milillo, Pietro	425
Minami, Takahiro	317
Mitchell, Andrew	221
Molch, Katrin	217
Monsorno, Roberto	205

Montironi, Marco	379
Montmory, Alain	1
Morbidelli, Roberto	309
Moreau, Agathe	225
Moreau, Yoann	344
Moreno, Laura	259
Morgado, Miguel	469
Morillon, Pascal	371
Morin, Christine	371
Morris, Edward P.	247
Mougnaud, Philippe	142, 461
Moyano, Gorka	481
Mueller, Norman	36
Muerth, Markus	461
Muller, Jan-Peter	59
Mulone, Angelo Fabio	309, 359, 379
Murphy, Kevin	221
Naeije, Marc	481
Naeimi, Vahid	40
Naemi, Vahid	9
Natali, Stefano	236
Nativi, Stefano	32, 98
Navarro, Victor	259
Neglia, Silvio Giuseppe	394
Neuhaus, Christoph	154
Neumann, Geoffrey	445
Ngo, Robert	279
Nico, Giovanni	375
Nicolini, Gianalfredo	359
Nielsen, Allan A.	126
Nieto, Sara	193
Nitti, Davide Oscar	473
Nonin, Philippe	146
Nosavan, Julien	75, 225, 279
Notarnicola, Claudia	205
Noval, Louis	5
Nunes, Paulo	275
Nutricato, Raffaele	473
Okori, Jimmy	469
Olaizola, Igor G.	433
Oliver, Simon	36
Orrù, Carla	414
Ortner, Mathias	146, 387
Osborne, Lisa	469
Ostermann, Frank	63
Ottavianelli, Giuseppe	211
Ottosen, Thor-Bjørn	402
Oukil, Youcef	398
Ouzounis, Georgios	255
Pailler, Frederic	106
Palacin, Hugo	5
Panem, Chantal	106
Papadimitriou, Constantinos	421
Papandrea, Enzo	56

Parra, Cyrille	321
Paulin, Mireille	134, 344
Pavon Carrasco, Francisco Jose	363
Pelich, Ramona	67
Pelloquin, Camille	259
Pepler, Sam	287
Percivall, George	83
Perissin, Daniele	425
Permana, Hans	243
Perrone, Loredana	363
Pesaresi, Martino	52
Petch, Geoffrey	402
Petrat, Johannes	29
Piergentili, Fabrizio	25
Pinto, Salvatore	233
Piscini, Alessandro	363
Pitcher, Heather	469
Plyer, Aurélien	437
Politi, Eirini	453
Politis, Panagiotis	52
Poupart, Emmanuel	344
Poupart, Erwann	134
Pritchard, Matt	287
Pryor, Matt	287
Purss, Matthew	32
Pödör, Zoltán	410
Quartulli, Marco	375, 433
Queiroz, Gilberto	48
Queyrut, Olivier	321
Ramapriyan, Hampapuram	221
Ramos, Jose Julio	110
Rapiński, Jacek	336
Rauste, Yrjö	239
Reato, Thomas	13
Reck, Christoph	313
Rees, Alan F.	465
Reid, Simon	469
Reinartz, Peter	293
Restano, Marco	477, 481
Riclet, François	106
Rilee, Michael L	90
Rivolta, Giancarlo	214, 340, 414
Robertson, Phil	449
Robinson, Ian	469
Roca, Mònica	481
Rodriguez Aseretto, Dario	52, 271
Rogier, François	437
Romeo, Antonio	233, 461
Roscher, Ribana	297, 301
Ross, Jonathon	36
Rouget, Matthieu	146
Rutakaza Maneno, Roger	279
Sabatino, Giovanni	67, 340, 477
Sabeur, Zoheir	445

Sabo, Filip	52
Sacramento, Paulo	79, 414
Sala Calero, Joan	247
Salgado, Jesus	193, 229
Samarelli, Sergio	473
Sanchez, Alber	48
Santillan, Daniel	236
Santoni, Fabio	25
Santoro, Francesca	363
Santoro, Mattia	98
Santoro, Maurizio	157
Santos, Rui	283
Scarrot, Rory	453
Scarth, Peter	130
Scherllin-Pirscher, Barbara	236
Schick, Michael	138, 325
Schindler, Konrad	293
Schmitt, Bruno	211
Schmullius, Christiane	157
Schrama, Ernst	481
Segur, Thierry	279
Seifert, Frank Martin	157, 239
Selle, Arnaud	279
Serio, Carmine	425
Shelestov, Andrii	122
Shimomura, Yuji	317
Shirkey, Gabriela	449
Shumilo, Leonid	122
Siemen, Stephan	44
Simoes, Rolf	48
Simonin, Mathieu	371
Siqueira, Andreia De Avila	36
Skjøth, Carsten Ambelas	402
Soille, Pierre	52, 71, 177, 267, 271
Solitto, Filomena	359
Soria-Frisch, Aureli	259
Soukup, Tomas	243
Souza, Ricardo Cartaxo Modesto	48
Spogli, Luca	363
Stamatiou, Kostas	255
Stefoudi, Dimitra	86
Stein, Thomas	185
Stilla, Uwe	293
Stoica, Adrian	414
Stoicescu, Miruna	138
Strobl, Christian	313
Strobl, Peter	32
Sudmanns, Martin	17, 173
Susino, Roberto	359
Suwala, Jason	461
Svaton, Vaclav	243
Sylos Labini, Giovanni	449
Syrris, Vasileios	52, 267, 271
Szantoi, Zoltan	32
Sánchez, Alber	441
Taillan, Christian	344

Taillan, Christophe	134
Tapete, Deodato	165, 425
Tavebard, Romain	371
Tedeschi, Cédric	371
Telloni, Daniele	359
Tergujeff, Renne	239
Terra-Homem, Miguel	453, 481
Thankapan, Medhavy	36
Thibaud, Balem	251
Tiede, Dirk	17, 173
Ties, Stephanie	469
Tincani, Lucrezia	469
Triebnig, Gerhard	236
Tripathi, Rachit	352
Trouvé-Peloux, Pauline	437
Tsarouchi, Gina	469
Tuan Sy, Le	40
Twele, André	313
Uebbing, Bernd	297
Uereyen, Soner	243
Vadaine, Rodolphe	371
Valentin, Bernard	142
Valeri, Massimo	56
Vallat, Claire	229
Valone, Elizabeth	469
van Bemmelen, Joost	98, 214, 239, 340
van der Velde, Marijn	367
Van Roey, Tom	461
van Zetten, Peter	239
Vanhuyse, Sabine	263
Vasalos, Georgios	421
Vasilev, Veselin	52, 271
Vasiliev, Vladimir	122
Vecchiato, Alberto	309
Ventimiglia, Florent	5
Ventrucci, Massimo	56
Ventura, Bartolomeo	205
Veres, Galina	445
Veskos, Paschalis	21
Vicente-Guijalba, Fernando	205
Viet, Phi	169
Vinhas, Lubia	48
Voidrot, Marie-Françoise	83
Volden, Espen	461
Volkman, Rouven	313
Wagemann, Julia	44
Wagner, Wolfgang	9, 40
Wegmüller, Urs	157
Wegner, Jan Dirk	293
Weiland, Nicolas	201
Wenzel, Susanne	297, 301
Wiesmann, Andreas	157
Williams, Alexa	469
Williams, Jamie	469

Wolf, Lothar	138
Wolff, Eléonore	263
Wu, Susie R.	449
Wunderle, Stefan	154
Yao, Wei	289, 293
Yeh, Laurent	189
Zeidler, Julian	243
Zeitouni, Karine	189
Zigna, Jean-Michel	445
Zinkiewicz, Daniel	336
Zinno, Ivana	209
Zurita Milla, Raul	63

Abstract

Big Data from Space refers to Earth and Space observation data collected by space-borne and ground-based sensors. Whether for Earth or Space observation, they qualify being called 'big data' given the sheer volume of sensed data (archived data reaching the exabyte scale), their high velocity (new data is acquired almost on a continuous basis and with an increasing rate), their variety (data is delivered by sensors acting over various frequencies of the electromagnetic spectrum in passive and active modes), as well as their veracity (sensed data is associated with uncertainty and accuracy measurements). Last but not least, the value of big data from space depends on our capacity to extract information and meaning from them.

The goal of the Big Data from Space conference is to bring together researchers, engineers, developers, and users in the area of Big Data from Space. It is co-organised by ESA, the Joint Research Centre (JRC) of the European Commission, and the European Union Satellite Centre (SatCen). The 2017 edition of the conference was hosted by CNES and held at the Pierre Baudis Convention Centre in Toulouse (France) from the 28th to the 30th of November 2017.

These proceedings consist of a collection of 126 short papers corresponding to the oral and poster presentations presented at the conference. They are organised in sections matching the order of the conference sessions followed by the contributions that were presented during the poster session, also organised by topics. They provide a snapshot of the current research activities, developments, and initiatives in Big Data from Space.

While a continued number of contributions are devoted to infrastructures and platforms enabling to exploit the value behind the volume, velocity, and variety of Big Data from Space, this third edition of the Big Data from Space conference shows a sharp increase of applications particularly related to large scale analysis including the temporal dimensions in view of better understanding the dynamics of the processes that are shaping our planet and our universe. Other new trends regard the information extraction using advanced machine learning techniques such as those based on deep learning and convolution neural networks. The development of new standards to ensure the interoperability of Big Data from Space is also gaining attention similarly to data cubes and multidimensional array representations. All these topics as well as other generic key aspects of big data are mirrored onto dedicated sections in these proceedings

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: <http://europea.eu/contact>

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: <http://europa.eu/contact>

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: <http://europa.eu>

EU publications

You can download or order free and priced EU publications from EU Bookshop at: <http://bookshop.europa.eu>. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see <http://europa.eu/contact>).

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.



EU Science Hub
ec.europa.eu/jrc



@EU_ScienceHub



EU Science Hub - Joint Research Centre



Joint Research Centre



EU Science Hub



Publications Office

doi:10.2760/383579

ISBN 978-92-79-73527-1