European Commission

JRC TECHNICAL REPORTS

The Global Conflict Risk Index (GCRI)
Manual for data management and product output

Version 5
Code documentation and methodology summary

GCRI
Global Conflict Risk Index

Martin Smidt
Luca Vernaccini
Peter Hachemer
Tom De Groeve

2016

EUR 27908 EN

Joint Research Centre

Contact information
Tom de Groeve
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 267, 21027 Ispra (VA), Italy
E-mail:tom.de-groeve@ec.europa.eu
Tel.: +39 0332786340

JRC Science Hub
https://ec.europa.eu/jrc

JRC100775

EUR 27908 EN

ISBN 978-92-79-58227-1 (PDF)

ISSN 1831-9424 (online)

doi:10.2788/705817 (online)

Printed in *Italy*

How to cite: Smidt, M., L. Vernaccini, P. Hachemer, T. De Groeve; The Global Conflict Risk Index (GCRI): Manual for data management and product output; EUR 27908 EN; doi:10.2788/705817

The Global Conflict Risk Index (GCRI)

Manual for data management and product output

Table of contents

Abstract

This technical report presents the methodology and code documentation for version 5 of the Global Conflict Risk Index. This release features changes to the data management system and imputation methods, and includes a new composite indicator.

A reliable data management system is presented, which combines the data from various sources and imputes missing data. The reproduction process for the two step regression model of previous versions.

This document focuses on the technical aspects of producing the dataset and statistical output. For details regarding the theoretical framework, please see the previous scientific report, doi 10.2788/184.

1. Introduction

The Global Conflist Risk Index (GCRI) is an early warning system designed to give policy makers a global risk assessment based on macro-economic factors.

Previous versions of the GCRI has incrementally developed a methodology for defining conflict, and the optimal regression model for predicting such conflicts. Two conflict dimensions were defined, and historical conflict data (Eck and Hulman, 2007; Sundberg et al., 2012; UCDP) was used to create a database of conflict events. Literature from the conflict science field was used to identify five theoretical risk areas. Within these, a further distinction was made between concepts, which were then represented with individual variables. The variables used are all relatively stable, in that little change is to be expected from year to year. A combination of a logistic and an OLS model is used to predict outbreaks of conflict, and to evaluate conflict risk. Raw data and the regression results are used to produce country profiles that present the risk and the background data used to calculate it.

This report presents the work done between September 2015 and February 2016, which has focused on quality control of the dataset, and on constructing a composite indicator.

A number of important features are added with this version of the GCRI. The construction of the dataset is now made more transparent and reproducible by the use of R scripts rather than the previous Excel based approach. The use of machine imputation for missing data has been replaced with a system where data is taken from either the closest known historical data, from regional averages, or from similar countries. The imputation is now included in the data construction phase, and so the data is now a single, complete dataset ready for statistical analysis.

The previous model validation work was repeated using the new data to verify that earlier findings still hold true. The old code was then cleaned of superfluous material, and the remaining code was adapted and simplified to make it easier to use. The code that comes with the current version is designed as a tool to produce the GCRI rather than to test multiple models.

In addition to the new data management system and output tools, a composite indicator component has been added. While the regression results are still used as the main basis for producing conflict predictions, the composite gives an added tool that for evaluating conflict potential.

While the work presented here shows great advances in reliability and reproducibility, there is still potential for improvements. The imputation method used for the data is now uniform among the indicators, and one improvement could be to implement specialized methods for the variables depending on their nature. There are also still many variables with very little data that rely heavily on these imputations. As such there is a possibility that these variables are introducing a great deal of noise.

The report is structured to go through the three main aspects of the GCRI in turn. First the dataset is introduced, including where the data is sourced, how the data is transformed, and finally a replication manual. Then the regression part of the GCRI is presented, going through the design and output, before continuing to the manual for the regression code. The composite indicator is then presented in the same manner, with a conceptual explanation and a replication manual.

2. Data

This section first introduces the data used, listing the variables, their sources, and how they are transformed. The second part is a manual for using the existing code, and for adding new data to the set within the existing framework. The data described here is version 5.0.2 of the dataset, and details may differ between versions.

2.1 Overview

The previous theoretical work established five distinct risk areas for conflict. Table 1 lists these five risk areas, their ten component concepts, and the 24 individual variables that are used to represent them. The data used remains the same as in previous versions, with the exception that the Conflict Trend variable has been dropped. Tables 2 and 3 give details about the variables' sources, while Table 4 present some descriptive statistics (except for the security variables).

Risk Area	Concept	Indicator
Political	Regime type	Regime Type
		Lack of Democracy
	Regime performance	Government Effectiveness
		Level of Repression
		Empowerment Rights
Social cohesion & Public security	Ethnic compilation	Ethnic Power Status (National Power)
		Ethnic Diversity (Subnational)
		Transnational Ethnic Bonds
	Public security & health	Corruption
		Homicide Rate
		Infant Mortality
Conflict prevalence	Current conflict situation	Recent Internal Conflict
		Neighbours with highly violent conflicts
	History of conflict	Years since highly violent conflict
Geography and Environment	Geographic challenge	Water Stress
		Oil Producer
		Structural Constraints
	Demographics	Population Size
		Youth Bulge
Economy	Development and distribution	GDP per capita
		Openness
		Income inequality
	Provisions and Employment	Food Insecurity
		Unemployment Rate

Table 1- Table of independent variables

Indicator	Source	Name of dataset	Name of original indicator(s)	URL
Regime type	Center for Systemic Peace	Polity IV Annual Time-Series, 1800-2014	Parcom, exrec	http://www.systemicpeace.org/inscrdata.html
Lack of democracy	Center for Systemic Peace	Polity IV Annual Time-Series, 1800-2014	Polity2	http://www.systemicpeace.org/inscrdata.html
Government effectiveness	World Bank	Worldwide Governance Indicators	Government Effectiveness: Estimate	http://databank.worldbank.org/data/reports.aspx?source=worldwide-governance-indicators
Level of repression	Political Terror Scale Project	PTS Data	Highest of the three indicators in the set	http://www.politicalterrorscale.org/Data/Download.html
Empowerment rights	CIRI Human Rights Data Project	CIRI Data	NEW_EMPINX	http://www.humanrightsdata.com/p/data-documentation.html
Ethnic compilation, NP	ETHZ	Ethnic Power Relations Core Dataset	Recoding of dataset, see variable page	http://www.icr.ethz.ch/data/epr
Ethnic compilation, SN	ETHZ	Ethnic Power Relations Core Dataset	Recoding of dataset, see variable page	http://www.icr.ethz.ch/data/epr
Transnational ethnic bonds	Center for International Development and Conflict management	(Minorities At Risk) MAR Quantitative	GC10	http://www.cidcm.umd.edu/mar/mar_data.asp
Corruption	World Bank	Worldwide Governance Indicators	Control of Corruption: Estimate	http://databank.worldbank.org/data/reports.aspx?source=worldwide-governance-indicators
Homicide rate	World Bank	World Development Indicators	Intentional homicides (per 100,000 people)	http://data.worldbank.org/indicator/VC.IHR.PSRC.P5
Infant mortality	World Bank	World Development Indicators	Mortality rate, under-5 (per 1,000 live births)	http://data.worldbank.org/indicator/SH.DYN.MORT
Recent internal conflict	UCDP/PRIO	Battle related deaths, Onesided violence, Non-state violence	Highest casualty estimates	http://www.pcr.uu.se/research/ucdp/datasets/
Neighbouring conflict	UCDP/PRIO	Battle related deaths, Onesided violence, Non-state violence	Highest casualty estimates	
Years since highly violent conflict	UCDP/PRIO	Armed Conflict Dataset	Conflicts of intensity level 2	

Table 2 - Variable sources

Indicator	Source	Name of dataset	Name of original indicator(s)	URL
Water stress	World Resources Institute	Aqueduct Country and River Basin Rankings (Raw country scores)	tdefm	http://www.wri.org/resources/data-sets/aqueduct-country-and-river-basin-rankings
Oil producer	World Bank	World Development Indicators	Fuel exports (% of merchandise exports)	http://data.worldbank.org/indicator/TX.VAL.FUEL.ZS.UN
Structural constraints	The Bertelsmann Stiftung	BTI 2014	Structural constrains (Q13.1)	http://www.bti-project.org/downloads/bti-2014/
Population size	UN Population Division	Annual population by single age - Both Sexes.	Sum of all ages	http://esa.un.org/unpd/wpp/Download/Standard/Interpolated/
Youthbulge	UN Population Division	Annual population by single age - Both Sexes.	Sum of ages 15-24 divided by sum of ages 25+	http://esa.un.org/unpd/wpp/Download/Standard/Interpolated/
GDP per capita	World Bank	World Development Indicators	GDP per capita, PPP (constant 2011 international $)	http://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD
Openness	World Bank	World Development Indicators	Foreign direct investment, net inflows (BoP, current US$)	http://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD
			Foreign direct investment, net inflows (% of GDP)	http://data.worldbank.org/indicator/BX.KLT.DINV.WD.GD.ZS
			Exports of goods and services (% of GDP)	http://data.worldbank.org/indicator/NE.EXP.GNFS.ZS
Income inequality	Frederick Solt	The Standardized World Income Inequality Database	Net inequality	https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/11992
Food insecurity	FAO	Food security indicators	Average dietary energy supply adequacy	http://www.fao.org/economic/ess/ess-fs/ess-fadata/en/
			Domestic food price index	
			Prevalence of undernourishment	
			Domestic food price volatility	
Unemployment	World Bank	World Development Indicators	Unemployment, total (% of total labor force) (modeled ILO estimate)	http://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
Conflict intensity	UCDP/PRIO	Battle related deaths, Onesided violence, Non-state violence	Highest casualty estimates	http://www.pcr.uu.se/research/ucdp/datasets/

Table 3 - Variable sources

Indicator	Original range	Min	Max	Transformation (Before rescaling)	Years covered	Missingness
Regime type	Parcomp 0 to 5	NA	NA	See variable details.	1989-2014	Palestine
	Exrec 1 to 8	NA	NA		1989-2014	Palestine
Lack of democracy	-10 to 10	None	None	None	1989-2014	Palestine and partial Lebanon
Government effectiveness	-2.5 to 2.5	None	None	None	1996, 1998,2000, 2002-2014	1319
Level of repression	1 to 5	None	None	None	1989-2014	29
Empowerment rights	0 to 14	None	None	None	1989-2011	627
Ethnic compilation, NP	NA	NA	NA	See variable details.	1989-2013	Event data, assumed complete
Ethnic compilation, SN	NA	NA	NA	See variable details.	1989-2013	Event data, assumed complete
Transnational ethnic bonds	0 to 5	None	None	None	1989-2006	3187
Corruption	-2.5 to 2.5	-2	2	None	1996, 1998,2000, 2002-2014	1306
Homicide rate	0.2 to 139.13	None	50	None	1995-2012	2382
Infant mortality	2.7 to 332.9	None	250	None	1989-2014	0
Water stress	1.9 to 4.44	None	None	None	NA	One observation for each country
Oil producer	0 to 99.79	None	None	Log	1989-2014	1140
Population size	303k to 1,369,435k	403k	268,337k	Log	1989-2014	0
Structural constraints	1 to 10	None	None	None	2006, 2008, 2010, 2012, 2014	2928
Youth bulge	0.14 to 0.40	None	None	None	1989-2014	0
GDP per capita	246.7 to 139456.8	500	100000	Log	1989-2014	310
Income inequality	0-1	None	None	None	1989-2014	1288
Food insecurity	Nourishment 5 to 80	None	35	None	1990-2014	
	Volatility 0 to 210.4	None	20	None	2000-2014	
	Diet 68 to 165	75	150	None	1990-2014	
	Price level 1 to 11.69	None	10	None	2000-2014	
Openness	-4.83e+10 to 3.02e+10	100k	15billion	Log	1989-2014	927
	-82.89 to 466.56	**1**	15	Log	1989-2014	1189
	0.18 to 230.27	3	200	Log	1989-2014	1028
Unemployment	0 to 39.3	1	None	Log	1989-2014	400

Table 4 - Variable details

2.2 Variables

All the independent variables are either rescaled from their original values to a 0 to 10 scale, or manually coded with values in this range. Before this transformation, the treatment differs. Table 4 presents a number of descriptive statistics for the data. The original range gives the minimum and maximum value on the variable before the data was transformed in any way. The min and max columns display the artificial minimum and maximum values that were enforced on the data. Units scoring above or below this are given the maximum or minimum scores. The transformation column shows which variables were log transformed before rescaling, and those who were subject to more complex coding. The years covered by the dataset, and the number of missing units are also noted. Where the missingness is only one or two countries, these are named.

2.2.1 Conflict Intensity

Both dependent variables are derived from UCDP/PRIO casualty data, taken from the Battle Related Deaths (BRD) (Sundberg, 2008), One Sided Violence (OSV), and Non-State Conflict (NSC) datasets. Depending on the nature of the conflict, they are designated as either National Power conflicts, or Subnational conflicts. The two dimensions are treated separately, and the GCRI scores on either dimension are independent of the other. The conflicts in each dataset are classified on an intensity scale from 0 to 3 (<25 deaths = 0, 25-499=1, 500-999=2, >1000=3). The conflicts from the three datasets are then combined into a single set, and aggregated to a country-year level with the maximum intensity and number of conflicts recorded. Each country-year is then classified on a scale from 0 to 10 based on the maximum intensity and the number of conflicts. This conflict intensity is the dependent variable used by the OLS model.

GCRI observed conflict intensity	Max intensity	No of conflicts
5	1	1
6	1	2
7	1	>2
8	2	1
8.5	2	2
9	2	>2
10	3	>=1

Table 5 - Intensity coding rules

This intensity is also used to create two dummy variables: Violent Conflict (VC) and Highly Violent Conflict (HVC). These are coded 1 for any unit with intensities of >=8 is coded a 1, and all others 0. These variables can be used for the logistic regression model. The dummy is also used to subset the units to be used in the OLS model. While the log model uses all units in the set, the OLS is run using only the units that are above the intensity threshold chosen for the logistic model.

The BRD dataset is split between conflict dimensions based on the type of conflict, and the incompatability recorded in the set. Internal and internationalized internal conflicts over governmental powers (Types 3 and 4, incompatibility 2) are coded as National Power conflicts. Extrasystemic, internal, and internationalized conflicts over territory are coded as Subnational conflicts. Other conflicts in the BRD dataset are not used by the GCRI.

Conflicts in the OSV dataset that involve a government actor are coded as Subnational conflicts. In countries where the conflicts are recorded with more than one location, the conflict is attributed to the country whose government is involved.

The conflicts in the NSC dataset are also coded as Subnational conflicts, and where more than one location is recorded the conflict is attributed to the country where the majority of the conflicting groups are situated.

2.2.2 Regime Type

The regime type indicator is designed to capture the effect of the democratic U-curve, where anocracies are seen as inherently less stable than autocracies and democracies (Hegre, 2001). The variable is constructed using Goldstone et al's (2010) methodology, which combines two components of the Polity IV project's Polity variable.

Countries are classified according to Table 6, which then decides which value it will be given on the regime type variable. Palestine is the only country without complete data, and it is coded as a partial autocracy.

	Competitiveness of Political Participation					
Executive Recruitment	**Repressed (0)**	**Suppressed (1)**	**Unregulated (2)**	**Factional (3)**	**Transitional (4)**	**Competitive (5)**
(1) Ascription						
(2) Ascription + Designation						
(3) Designation						
(4) Self-Selection						
(5) Transition from Self-Selection						
(6) Ascription + Election						
(7) Transitional or Restricted Election						
(8) Competitive Election						

Based on POLITY IV scales for Executive Recruitment (EXREC) and Competitiveness of Political Participation (PARCOMP).

White = Full Autocracy Light grey = Partial Autocracy Dark grey = Partial Democracy

Very dark grey = Partial Democracy w/factionalism Black = Full Democracy

Table 6 - Regime type coding rules from Goldstone et al. (2010)

The different regime types are then scored according to Table 7, which also includes the Transition and Foreign Intervention categories. These are used for the countries who are not given scores on the EXREC and PARCOMP variables because of either a period of transition or because of an ongoing foreign intervention.

Type:	Full autocracy	Partial autocracy	Partial democracy w/factionalism	Partial democracy	Full democracy	Transition	Foreign Intervention
Score:	1.1	4	10	3.9	1	4.49	10

Table 7 - Regime type scores

2.2.3 Lack of Democracy

This variable is simply the polity2 variable from the same Polity IV dataset as the regime type variable, but rescaled from it's original -10 to 10 to a 0 to 10 scale. Palestine is the only country with no data, and it is coded as a -2. Lebanon is missing data from 1990 to 2004, and these years are coded as 4.

2.2.4 Government effectiveness

Government effectiveness is taken from the World Bank's Worldwide Governance Indicators database. The original data is an estimate of government effectiveness based on a range of factors, such as perceptions of the quality of public services, civil services, and these service's independence from political control (Kaufmann et al. 2010). Countries are scored in units of a normal distribution, so the range of the original variable is approximately -2.5 to 2.5. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.5 Level of Repression

Repressing media and civil liberties is a way of suppressing unrest, and by hindering the use of peaceful channels the regime pushes the opposition towards other means. High levels of repression has been found to correllate with conflict by both Fox (2004) and Regan and Norton (2005).

The repression variable is taken from data gathered by the Political Terror Scale Project. The dataset contains three variables, each coded from 1 to 5, based on data from Amnesty, Human Rights Watch, and the US State Department. The highest score of the three variables is used as the repression variable. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.6 Government effectiveness

Data is taken from the NEW_EMPINX variable from the CIRI Human Rigts Data Project. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.7 Ethnic Power Status

The data is taken from ETH Zürich's Ethnic Power Relations dataset. The data is originally in ethnic group-country-period format, with the status of each group recorded in a categorical variable. These categories are then divided into two groups depending on whether they are seen as being included in the political process, or excluded from it. Transitions from one status to another are then coded for each group-country, depending on the risk such a transition poses to stability. For example, a group going from being a dominant group to any excluded status will earn the country-year a score of 9. The score is given only to the country-year where the change occurs, which is taken to be the first year of the new status. For full details on the transitions see the R script for ETHNIC_NP.

STATUS	*SCORE*
State collapse	10
Self-exclusion	9
Regional autonomy	7
Discriminated	5
Junior/Senior partner	5

2.2.8 Ethnic Diversity

Dominant/Monopoly	1
Powerless/Irrelevant	1

The same data as the previous variable is used, except here the data is merely copied out to country-year format. The groups are then given values depending on their status, and each country year is coded using the highest score of all ethnic groups present in that year. A maximum score is given if there is a state collapse registered that year, and groups that are self-excluding are given a score of 9. Groups that want regional autonomy, and aren't powerless or irrelevant, are coded as 7. Discriminated groups and groups that are part of a partnership where power is shared are scored 5 when they do not seek regional autonomy. Dominant gorups and groups with a monopoly on power are scored 1. The powerless and irrelevant groups are always scored 1.

2.2.9 Transnational Ethnic Bonds

The original variable has poor coverage, and so a lot of data has to be imputed. 38 countries have no data, and so other countries or regional averages were used. See the R scripts for details on which countries are missing and which data is used. Apart from imputation, the variable is rescaled, with no further transformations or limits imposed.

2.2.10 Corruption

The corruption indicator measures public perception of government representatives' use of their office for personal gain. The source indicator is given as units of a standard normal distribution, giving the relative score of countries with regards to each other (Kaufmann et al. 2010).

Due to outliers, a minimum and maximum value of 2 was enforced. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.11 Homicide Rate

Any added effect of increased values is believed to diminish at extreme values, but not following a log curve. A maximum value of 50 was enforced. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.12 Infant Mortality Rate

A measurement of public health and welfare, where low levels represent poor living conditions that make armed rebellion a more appealing choice (Collier & Hoeffler 2004).

Any added effect of increased values is believed to be negligible at extreme values, but not following a log curve. A maximum value of 250 was enforced. The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.13 Recent Internal Conflict

Conflict history is among the more powerful predictors of conflict (Collier et al., 2008; Hegre et al., 2013). Conflict history is not only important because the old conflict might continue, but because of the presence of weapons and experience fighters makes it easier for violence to continue or restart. As the dependent variable is conflict in the future, the recent internal conflict variable records the current conflict situation.

This variable is coded as the highest conflict intensity of the two dimensions for the year in question.

2.2.14 Neighbouring Conflict

Spillover effects between countries are modelled by the inclusion of a neighbourhood variable. Conflicts can spread from country to country through several mechanisms. A direct mechanism can be armed groups crossing international borders and destabilizing

the neighboring countries. As countries are often highly linked with their neighbours, both economically and through other means, neighbouring conflicts can also cause unrest through their disruption of the economy, through refugee flows, or through a flow of illegal weapons.

This variable is the highest conflict intensity of the two dimensions in a neighboring country. Neighbours are defined using the cshapes package for R, which calculates the minimum distance between any two countries. A country's neighbours are any countries whose nearest point is less than a kilometer away from their borders. The highest score among the neighbours is used as the score.

2.2.15 Years Since Highly Violent Conflict

Another conflict history variable that measures the time since a major conflict event. This

This variable is an inverted count of years since the last HVC in the country, with a maximum of 10 years. As data is needed from earlier than what is covered in the datasets used to calculate the, the UCDP/PRIO Armed Conflict Database (Gleditsch et al, 2002) is used. This dataset has coverage back to ten years before the other variables. It only covers battle related casualties, but the definition of a highly violent conflict is the same as used by the GCRI conflict intensity (1000+ casualties).

2.2.16 Water Stress

The variable is the *tdefm* variable of the World Resource Institute's Aqueduct Country and River Basin Rankings dataset (The variable is one of the raw components, not one of their indexes). The data is not a time series, but judgement of the country based on time series data. All countries are given one value, which is used for the entire period. The variable is rescaled, with no further transformations or limits imposed.

2.2.17 Oil Exporter

The variable is the proportion of a country's GDP that is export of fossil fuels. The variable is log transformed, imputed and rescaled, with no further limits imposed.

2.2.18 Structural Constraints

The original variable is the Structural Constraints (Q13.1) variable from the Bertelsmann Transformation Index. Data from 2003 is not used as the variable was not recorded independently of its sub-index at that point. The variable is imputed and rescaled, with no further transformations or limits imposed. Note that a high number of countries were imputed using regional averages/similar countries.

2.2.19 Population Size

Demographics, especially population size, is amon the most commonly used indicators in conflict research. Population size is among the most robust indicators, even though it's theoretical link remains unclear (Fearon & Laitin, 2003; Collier & Hoeffler, 2004; Hegre & Sambanis, 2006).

The total population, log transformed. The variable has complete data, and so no imputation is needed. Before log transformation, minimum and maximum values are enforced on outliers.

2.2.20 Youth Bulge

Another demographic aspect found to correlate with conflict is youth bulges (Gurr et al. 1999). The large amount of young people compared to the rest of the population is theorized to provide easier recruitment opportunities to rebel groups.

The number of inhabitants between 15 and 24 divided by the number of inhabitants 25 and over. As with the population variable, the data is complete and no imputation is needed. No further transformations or limits are imposed.

2.2.21 GDP Per Capita

GDP per capita is consistently linked with conflict in literature (Hegre & Sambanis 2006). A low level of income makes it easier to recruit solders to rebel organisations, as making a living from normal work becomes harder (Collier & Hoeffler 2004).

The GDP PPP per capita in 2011 USD, taken from the World Bank's World Development Indicators. Outliers were forced to a minimum value of 500, and a maximum of 100,000, before log transformation was applied.

2.2.22 Openness

The variable is constructed using three economic indicators from the World Bank's Word Development Indicators: Foreign Direct Investment (FDI) in net dollars, in percentage of GDP, and exports as part of the GDP. Min and max values for outliers are enforced for all three, imputations are made, and they are all log transformed. They are then rescaled before they are combined into the final indicator. This is done by first averaging the two FDI variables, then averaging the result with the export variable.

2.2.23 Income Inequality

The data source for the predictor is the Standardized World Inequality Inequality Database's Net Inequality variable. This is designed to be comparable both between countries and over time, and to have global coverage.

The variable is imputed and rescaled, with no further transformations or limits imposed.

2.2.24 Food Insecurity

Price level, price volatility, dietary requirements, and nourishment are combined to create an indicator of food insecurity. The four component variables are all imputed, and rescaled before being merged to a single indicator. The indicator is created by a weighted average of the price level and the price volatility, which is again averaged with the other two variables.

*mean(DIET,NOURISHMENT,(0.8*PRICE LEVEL)+(.2*PRICE VOLATILITY))*

2.2.25 Unemployment Rate

Definitions vary, but usually the percentage of the population that is working age and unemployed. The data is taken from the World Bank's database, which contains data from multiple sources. Sources vary between countries, with some using international assessments, and others using numbers reported by national offices.

The variable is imputed and rescaled, with no further transformations or limits imposed.

In the finished dataset, all the dependent variables have a range of 0 to 10 and 0 missing. Figure 1 is a correlation matrix of the variables, showing that some variables are highly correlated.

	YRS_HVC	CON_INT	CON_NB	CORRUPT	DISPER	ECON_ISO	EMPOWER	ETHNIC_NP	ETHNIC_SN	FOOD	FUEL_EXP	GDP_CAP	GOV_EFF	HOMIC	INEQ_SWIID	MORT	POP	REG_P2	REG_U	REPRESS	STRUCT	UNEMP	WATER	YOUTHBBOTH
YRS_HVC																								
CON_INT	0.67																							
CON_NB	0.19	0.26																						
CORRUPT	0.23	0.29	0.34																					
DISPER	0.20	0.31	0.36	0.25																				
ECON_ISO	0.22	0.26	0.21	0.21	0.02																			
EMPOWER	0.14	0.20	0.34	0.37	0.16	0.09																		
ETHNIC_NP	0.08	0.11	0.06	0.09	0.06	0.05	0.01																	
ETHNIC_SN	0.11	0.20	0.10	0.08	0.16	0.04	0.17	0.06																
FOOD	0.22	0.22	0.27	0.61	0.07	0.31	0.15	0.09	0.05															
FUEL_EXP	0.06	0.06	0.09	0.09	0.03	-0.18	0.31	-0.01	0.02	-0.21														
GDP_CAP	0.20	0.24	0.31	0.61	0.13	0.38	0.10	0.10	0.10	0.79	-0.36													
GOV_EFF	0.23	0.26	0.31	0.90	0.16	0.29	0.38	0.10	0.12	0.66	0.00	0.69												
HOMIC	0.11	0.12	0.07	0.40	0.13	0.05	-0.19	0.04	-0.19	0.41	-0.14	0.36	0.39											
INEQ_SWIID	-0.06	-0.06	-0.09	0.11	-0.06	0.02	-0.16	0.01	-0.03	0.30	-0.16	0.18	0.12	0.45										
MORT	0.20	0.26	0.37	0.64	0.11	0.34	0.16	0.11	0.14	0.81	-0.14	0.81	0.69	0.44	0.29									
POP	0.23	0.35	0.27	0.08	0.39	0.18	0.17	0.04	-0.02	-0.06	0.15	0.03	-0.05	-0.04	-0.13	-0.03								
REG_P2	0.08	0.07	0.26	0.32	0.00	0.13	0.77	-0.03	0.09	0.22	0.30	0.13	0.37	-0.16	-0.07	0.29	-0.06							
REG_U	0.17	0.19	0.05	0.32	0.21	0.01	-0.10	0.07	0.09	0.20	-0.16	0.25	0.28	0.19	0.08	0.18	0.00	-0.23						
REPRESS	0.50	0.62	0.35	0.55	0.35	0.28	0.43	0.09	0.14	0.34	0.11	0.37	0.51	0.22	0.05	0.36	0.42	0.26	0.23					
STRUCT	0.29	0.33	0.37	0.71	0.20	0.30	0.29	0.11	0.13	0.76	-0.12	0.80	0.77	0.39	0.19	0.80	0.05	0.30	0.25	0.48				
UNEMP	0.05	0.02	0.03	0.07	0.05	-0.09	-0.03	-0.03	0.02	-0.09	0.06	-0.06	0.08	0.13	0.04	0.00	-0.16	-0.06	0.10	0.05	-0.01			
WATER	0.04	0.02	0.01	0.01	-0.18	-0.02	0.24	-0.02	0.02	-0.01	0.06	-0.11	0.01	-0.22	-0.05	-0.04	-0.04	0.23	-0.05	0.06	-0.05	0.16		
YOUTHBBOTH	0.15	0.20	0.33	0.53	0.07	0.26	0.17	0.07	0.10	0.72	-0.15	0.69	0.60	0.39	0.31	0.83	-0.08	0.27	0.18	0.33	0.69	0.06	0.14	

Figure 1 - Correlation matrix

2.3 Replication manual

This section contains instructions on how to use the existing code framework to replicate the dataset. The dataset comes with all the unedited source data, and the documentation needed for replication. The data was transformed and combined using the statistical programming language R. To be able to reproduce the results you will need an installation of R, which is available for free here: https://cran.rstudio.com/

It is also recommended that you install an Integrated Development Environment (IDE), such as RStudio, also available as a free download here:

https://www.rstudio.com/products/rstudio/download/

Even without an R installation, you can view the code by simply opening the .R files in a text editor like Word or Notepad. The code is annotated extensively, and many parts can easily be followed even without knowledge of the R language.

The system is portable, not in the inter-system sense, but in that after changing one line of code, it can be run on any computer. It is however critical that *all original files are kept and the subfolders are left as they are.* None of the original files can be renamed or moved within the folder. The IDEP_VARIABLE folder must also be kept free from any files other than the original variable scripts (except if you wish to include further variables, and have written a compatible script).

2.3.1 Contents of the replication package

The main folder should contain the following:

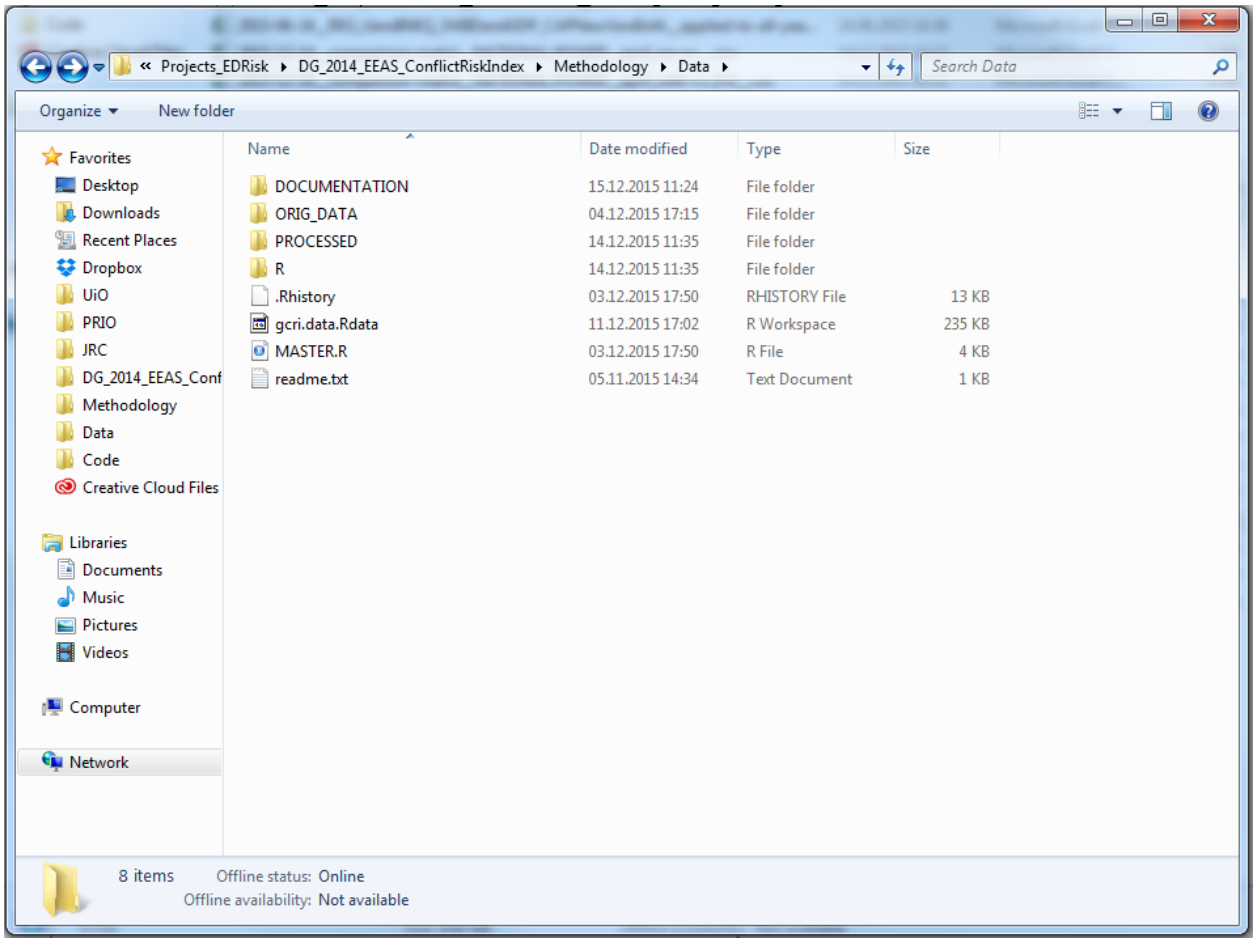


Figure 2 - Contents of the GCRI data folder

gcri.data.Rdata – Datafile containing the complete dataset. Is created by MASTER.R

MASTER.R – The main R script used to create the dataset. This script calls on a number of subscripts found in the R folder, which in turn process the variables and perform other support operations.

readme.txt – Short .txt containing a short briefing.

DOCUMENTATION - Folder containing this file, and the data presentation, along with supporting graphics and tables.

ORIG_DATA – Folder containing the raw data as it was downloaded from the respective sources, untouched apart from being renamed. Each variable has its own data file (apart from POP, which uses YOUTHBBULGE), in one of the following formats: .Rdata, .xlsx, .xls, .dta, or .csv.

PROCESSED – Folder containing an .Rdata file for each variable, which contains the individual variables in country-year format and with imputed data.

R – Contains all the scripts called by **MASTER.R**. Contains one subfolder for independent variables, and one for the dependent variable (conflict intensity) and the autoregressive variables (conflict history and neighboring conflict). It should also contain the following files:

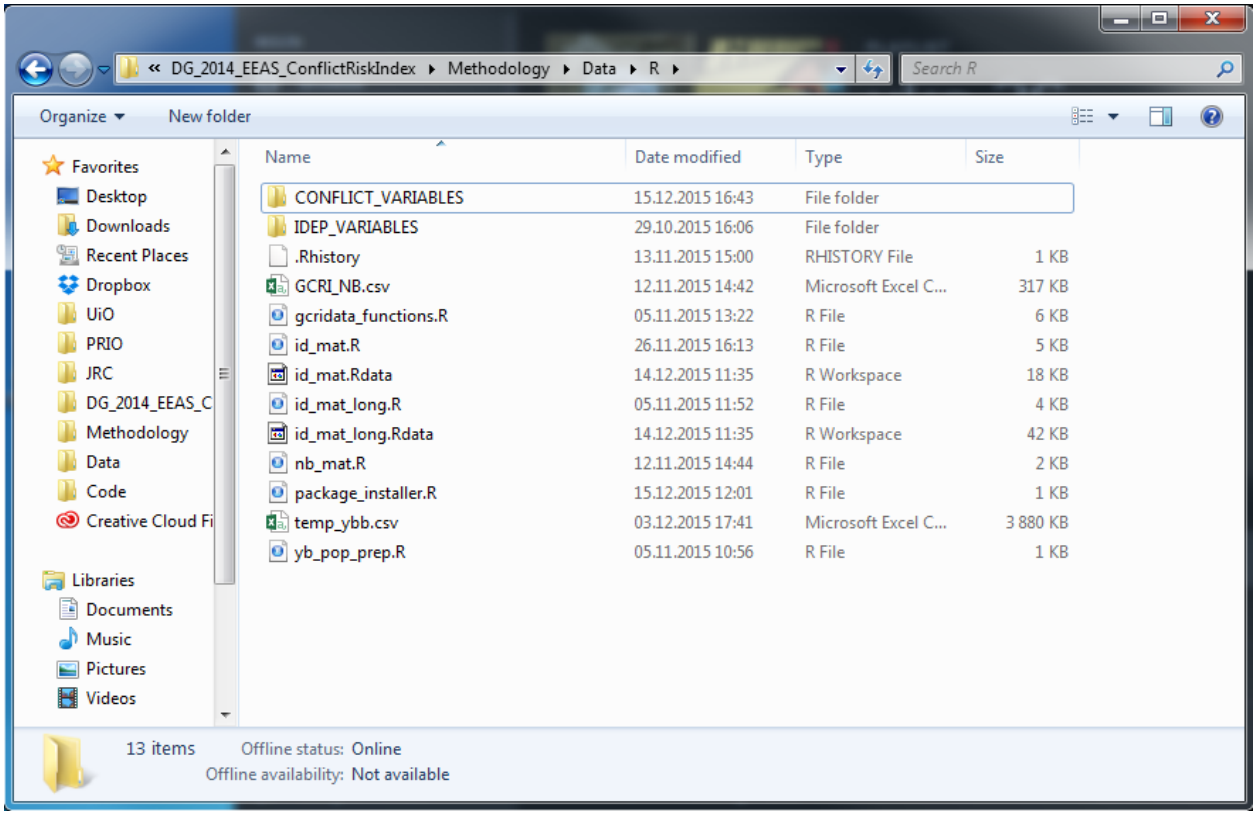


Figure 3- Contents of the R folder

GCRI_NB.csv – Dataset that identifies neighbors of each countries for all years. Created by nb_mat.R

gcridata_functions.R – Script containing custom functions. Sourced early in MASTER.R.

id_mat.R(data) – Script that creates **id_mat.Rdata**, a datafile that contains all the country-years included in the dataset, as well as a range of ID codes for each country. This file is used to merge datasets using many different ID coding systems, and serves as a register of all units that are wanted in the set.

id_mat_long.R(data) – Same as id_mat but going further back. Used for conflict history beyond the time span of the dataset.

nb_mat.R – Creates GCRI_NB.csv using the cshapes package. Takes ages to run.

temp_ybb.csv – Datafile created early in the master script to facilitate the processing of YOUTHBBOTH and POP. It is created by the **yb_pop_prep.R** script, and is a processed version of the YOUTHBULGE_BOTH_ORIG.xlsx file which is raw population data from the UN.

yb_pop_prep.R – Script that create temp_ybb.csv. The original file is quite large, and takes a long time to read into R, so this script downsizes and stores in a format that can be read more quickly (as the data is needed for more than variable, this saves some time later).

CONFLICT_VARIABLES – Folder containing the scripts used to create the conflict based predictors, and the dependent variables.

IDEP_VARIABLES – Folder containing the scripts used to create the independent variables. *It is crucial that this folder is kept free from any files other*

<u>than R scripts containing code that results in datasets that conforms with the others</u> (has complete data for all units in the id_mat, and can be merged by variables named ISO3C and YEAR).

2.3.2 Walkthrough of the code

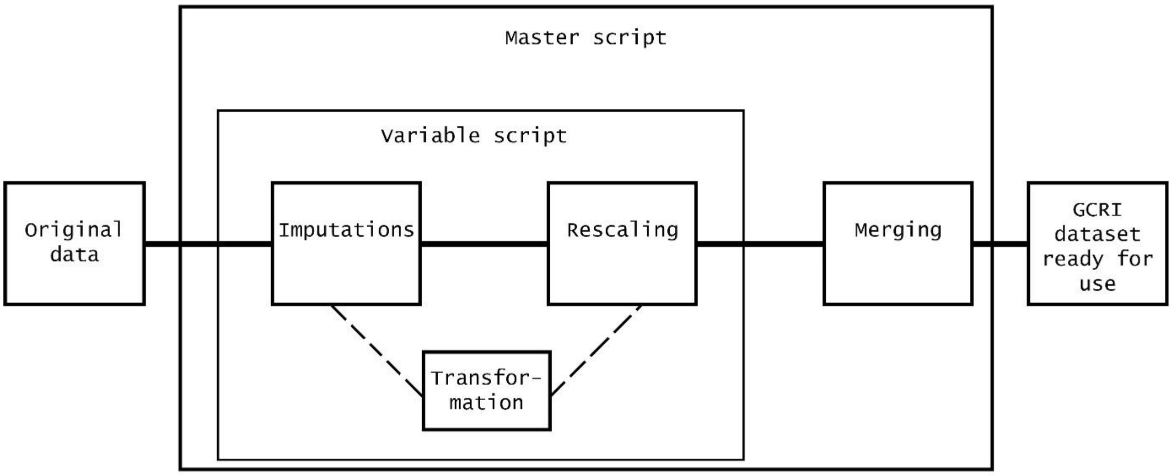


Figure 4 - Flowchart of the data management

MASTER.R

The first action is to set the working directory. Before running anything, you need to change the setwd(<path>) to where you have the folder on your own system. **This is the only hardcoded path, and all subsequent changes in working directories are done relative to this one.** Everything works as long as all subfolders are not renamed or moved.

```
#§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§#
##                  THIS IS THE MASTER WORKING DIRECTORY                      ##
##   MAKE SURE THIS LEADS TO THE PARENT FOLDER CONTAINING THE PROJECT        ##
setwd('C:/Users/ArthurDent/Data/GlobalConflictRiskIndex')
#§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§§#
```

The first chunk of code loads packages containing all necessary functions. A line is included, but commented out, which sources the **package_installer.R** script.

```
#source("R/package_installer.R")
```

This script is just an assistant that installs all the packages so you don't have to do it manually. The first time you run the MASTER.R you can uncomment this, or run it manually before starting. After having done it once, it is unnecessary for subsequent runs.

The next chunk prepares auxiliary functions, matrices, and datasets used in the main processing.

First, and commented out by default, **yb_pop_prep.R** is sourced. This script reads the YOUTHBULGE_BOTH_ORIG file found in the ORIGINAL_DATA folder, and converts it to a smaller .csv. The data in this file is used for both YOUTHBBOTH and POP variables, but reading it takes some time. Reading the semi-processed .csv file takes far less time, so preparing this file once and then skipping it will save time when running the entire MASTER script.

```
47  # Source a script that preps UN population data (for some reason this didn't work in the loop)
48  # This takes a while due to the huge dataset involved. Skip if this is not the first time you run.
49  # 22 warnings about NA introduced will appear. They are a good sign.
50  #source("R/yb_pop_prep.R")
51  # Creates "temp_ybb.csv" in the ORIG_DATA folder that will be used by the YBB and POP variable scripts.
```

id_mat.R and **id_mat_long.R** create matrices that contain a range of ID codes for all countries in the dataset. id_mat is also the guide for which country-years are to be included, and for which missing data should be imputed when not already in the data. The individual variable processing scripts all use this matrix, merging it with the original data. id_mat_long contains the same countries, but with years going back further than 1989. This one is used for conflict history only.

```
53  # Source scripts that create matrices containing the countries and country years that are to be kept.
54  source("R/id_mat.R")
55  source("R/id_mat_long.R")
```

nb_mat.R uses the cshapes package to create a dataset which identifies the neighbors of all countries for each year. The definition of a neighbor is taken as countries sharing land borders. This script takes many hours to run, so the output is included and the command commented out by default. Stores the grid as GCRI_NB.csv in the R folder.

```
56  #source("R/nb_mat.R") # Takes hours. Output is included, so run this only if you want to redefine what constitues a neighbour.
```

gcridata_functions.R contains custom functions used to transform and impute data.

```
57  # Load necessary custom functions
58  source("R/gcridata_functions.R")
```

first.last.impute repeats the first and last data points in a vector if there are missing values at the beginning or end. *repeat.previous.impute* fills in missing gaps in vectors by repeating the last value before the missing period. *flp.impute* is a support function that calls the other two in the correct order and returns the data in the proper format. *flp.impute* requires 3 arguments: the data frame containing your data, the column number of an identifying variable (ISO/COUNTRY/GWNO), and the column number of the variable you need imputed. It will return the same frame but with no missing for all countries that have at least one data point for that variable. Countries that are completely missing must have values assigned manually.

```
data<-flp.impute(data,1,12)
```

Using column numbers

```
data<-flp.impute(data,
                 which(colnames(data)=="ISO3C"),
                 which(colnames(data)=="CORRUPT"))
```

Using variable names

	1989		1991	1992	1993	1994	..	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Original			4.2	4.1	4.3				5.8	5.9	6.3			7	7.5	7.9	9.3	10	11.3			
Step 1	4.2	4.2	4.2	4.1	4.3				5.8	5.9	6.3			7	7.5	7.9	9.3	10	11.3			
Step 2	4.2	4.2	4.2	4.1	4.3				5.8	5.9	6.3			7	7.5	7.9	9.3	10	11.3	11.3	11.3	11.3
Step 3	4.2	4.2	4.2	4.1	4.3	4.3	..	4.3	5.8	5.9	6.3	6.3	6.3	7	7.5	7.9	9.3	10	11.3	11.3	11.3	11.3
Output	4.2	4.2	4.2	4.1	4.3	4.3	..	4.3	5.8	5.9	6.3	6.3	6.3	7	7.5	7.9	9.3	10	11.3	11.3	11.3	11.3

The way the flp.impute function completes data. Countries with no data are given the scores of similar countries or regional averages.

years.since.vector and *years.since* are used to calculate the time since a conflict occurred. The former does the actual work, while the latter is a shell function that can be applied to datasets when supplied with an id column and target column.

scale010 rescales data from whatever range it has to having a minimum value of 0 and a maximum of 10. The direction argument can be set to 0

or 1 depending on whether or not the target variable should also be inverted (i.e. the original max should be a 10 or a 0, and vice versa for the original min).

```
data$UNEMP <- scale010(data$UNEMP,0)
```

Everything is now ready to start loading variables. The next chunk sets the working directory to the one containing all the individual variable scripts, and then creates a list of all these.

```
#### WD change and file list ####
# Set WD to the folder containing independent variable scripts to retrieve list of vars to be processed.
setwd(idep_wd)
variable_list <- list.files()
```

This folder must be kept clean of any other files than the variable scripts, or the MASTER script will attempt to run these as well. New variables can be added by creating a new script and placing it in this folder. The script must be able to independently produce a dataframe named "data", which again must contain only three variables: ISO3C, YEAR, and the independent variable. The data output should contain only the country-years that are in the id_mat, and should be without any missing values.

The variable loop chunk sources the scripts from the variable list one by one, merging the resulting data frames together into a single frame.

```
for (variable in variable_list) {
  setwd(idep_wd)

  cat(paste("Sourcing variable script:",variable,"
            "))
  flush.console()

  source(variable)  # Source the individual script, retrieving a set named "data"

  if (!exists("ideps")) {
    # Create a dataset in which to combine the variables
    ideps <- as.data.frame(data)
    # send text message to the console
    cat(paste("variable '",colnames(ideps)[1],"' created and merged to set.
              "))
    flush.console()
  }else if (exists("ideps")) {
    # Add the rest of the variables to the set
    ideps <- merge(ideps, data, by = c("ISO3C","YEAR"), all.x = T)
    # send text message to the console
    cat(
      paste(
        "Variable '",colnames(ideps)[ncol(ideps)],"' created and merged to set. ",
        ncol(ideps) - 2,"predictors and ",nrow(ideps)," units now in the set.
        "
      )
      )
    flush.console()
  }
  rm(data) #clean
}
```

The loop goes through the list of variables in the IDEP_VARIABLES folder and executes all the scripts found there. *These scripts must always results in the creation of a dataframe named "data" that contains the new predictor, a variable named "ISO3C" containing the three letter "ISO" country code, and a variable named "YEAR" containing the year of the observation.* If they don't contain the two identifiers R will not be able merge it with the other data, and the script will stop.

The conflict variables are kept out of the main loop, and are called individually.

```
#### Conflict variables ####
setwd('R/CONFLICT_VARIABLES')
# Intensities based on PRIO/UCDP
source("CONFLICT_INTENSITY.R")
# YRS_HVC
source("YEARS_SINCE_CONFLICT.R")
# Y4 intensities
source("CON_LAG.R")
# Neighbouring conflicts
source("CONFLICT_NB.R")
```

CONFLICT_INTENSITY.R calculates conflict intensities based on casualty data from the three UCDP/PRIO datasets that contain such numbers (BRD, OSV, NSV). The deaths are credited to one of two conflict types, and conflict intensities are calculated based on these. Maximum intensity and number of ongoing conflicts for each year are used to set a GCRI conflict intensity.

YEARS_SINCE_CONFLICT.R calculates the time since the last highly violent conflict (HVC). For this there is no need for detailed casualty data, and so the UCDP/PRIO ACD can be used. This goes back further than the other three, and so data from 1979 is taken from here. The script calculates calculates how many years have passed since a violent conflict has occurred, up to a max of 10 years. This is then inverted so that countries with no conflict in the previous 10 years score 0, and a country with a violent conflict the previous year scores a 10.

CON_LAG.R creates the dependent variable actually used in the regression. One variable is created for each dimension, consisting of the maximum conflict intensity over the following 4 years of the country-year in question.

CONFLICT_NB.R creates the CON_NB variables by checking for conflicts in neighboring countries. Using **the GCRI_NB.csv** data created by **nb_mat.R,** each unit is assigned the highest intensity among its neighbors each year.

Following this, the last chunks remove supporting variables that are no longer needed, and merge the two sets of variables into one complete set that is compatible with previous GCRI versions. The workspace is cleaned, and the dataset is stored. As it is, the dataset is automatically stored as an .RData file, but it can easily be exported to Excel/STATA using the following commands:

Excel:
write.xlsx(gcri.data, "GCRI.DATA_v5.0.2.xlsx")

Stata:
write.dta(gcri.data, "GCRI.DATA_v5.0.2.dta")

As it stands, the dataset contains complete data for all 138 countries, for all the years they are defined to have existed in their current political form in the 1989 to 2014 period.

2.3.3 Adding new data

This section explains how new data can be added, either in the form of new releases of old data, or in the form of new variables. Adding data is relatively simple as long as the new versions of old data comes in the same format as before, in which case new years can be added almost automatically. Adding new variables takes more work, but only

requires you to be able to convert data to a country-year format, merge it with the id_mat, and apply the impute function (and fill in any remaining missing).

2.3.3.1 Updating with extra years

First of all, go to the id_mat.R and id_mat_long.R, and find the line where there id_mat(_long) is subset to cover a certain period (line 83 and 77 at the moment, may change. Code is commented so it should be easy to find). At the moment id_mat is set to 1989 to 2014, and id_mat_long to 1855 to 2014. Replace 2014 with 2015/2016 (or whatever year you want), save the script and run it so that a new .Rdata file is stored. This allows the individual variable files to keep data for the extra years, and to impute it from older data should it be missing. Because of some shortcuts in some of the variable scripts, they are manually limited to data newer than 1989, so changes to include older data will need extensive work.

Unfortunately, not all variable scripts are written in a way that makes them able to automatically handle new years. The data on structural weakness, STRUCT, comes in a format that makes automation problematic. If new data is to be included from this source, then you will have to manually script its inclusion (at the beginning of the STRUCT.R script you will need to read in the new sheet manually and add it to the list of other years that are merged).

All other variables should automatically include any newer data as long as the source file format remains the same (but with additional years added). STRUCT will also have data for all years in the id_mat even if the time range is expanded, but this will be copied from the 2014 data.

To include new data, download the new set from its source. ***Verify that the format is the same as the previous data from that variable, if it is not the script will not be able to read it.*** This means that it must not only be the same file format (.CSV/.XLSX/.RDATA and so on), but also be exactly the same in the grid system in the file. Column headers must start on the same row as before, data variables must start on the same columns, and so on. For most of the data this should not be a problem, as they either release data with new years as an additional column at the right hand end of the document, or as new country-year units.

If the new data is in the same format as before, simply delete the old file (or move to a backup folder), and rename the newly downloaded file to the name of the variable ("POP_ORIG.xlsx"). When you run the MASTER.R again, the new data should automatically be included (given that the id_mat.R script also has been edited to include the new year).

If you only edit the id_mat.R script to include new years, it will simply impute data for this year for any variables without new data for that year.

2.3.3.2 Adding/removing countries

At the start of the id_mat.R and id_mat_long.R there is a list of ISO3 character codes.

```
# List of 138 nations of interest (Non-EU and pop>500k)
isolist <- c("AFG", "AGO", "ALB", "ARE", "ARG", "ARM", "AUS", "AZE", "BDI",
"BEN", "BFA", "BGD", "BHR", "BIH", "BLR", "BOL", "BRA", "BTN",
"BWA", "CAF", "CAN", "CHE", "CHL", "CHN", "CIV", "CMR", "COD",
"COG", "COL", "COM", "CRI", "CUB", "DJI", "DOM", "DZA", "ECU",
"EGY", "ERI", "ETH", "FJI", "GAB", "GEO", "GHA", "GIN", "GMB",
"GNB", "GNQ", "GTM", "GUY", "HND", "HTI", "IDN", "IND", "IRN",
"IRQ", "ISR", "JAM", "JOR", "JPN", "KAZ", "KEN", "KGZ", "KHM",
"KOR", "KWT", "LAO", "LBN", "LBR", "LBY", "LKA", "LSO", "MAR",
"MDA", "MDG", "MEX", "MKD", "MLI", "MMR", "MNE", "MNG", "MOZ",
"MRT", "MUS", "MWI", "MYS", "NAM", "NER", "NGA", "NIC", "NOR",
"NPL", "NZL", "OMN", "PAK", "PAN", "PER", "PHL", "PNG", "PRK",
"PRY", "PSE", "QAT", "RUS", "RWA" ,"SAU", "SDN", "SEN", "SGP",
"SLB", "SLE", "SLV", "SOM", "SRB", "SSD", "SUR", "SWZ", "SYR",
"TCD", "TGO", "THA", "TJK", "TKM", "TLS", "TTO", "TUN", "TUR",
"TZA", "UGA", "UKR", "URY", "USA", "UZB", "VEN", "VNM", "YEM",
"ZAF", "ZMB", "ZWE")
```

This list decides which countries will be included in the set. If you add or remove countries this may lead to issues with the current code. Removing certain countries could create problems with scripts that copy country specific data rather than averaging when imputing. Adding countries might lead to other issues. If the country has no data on a variable, the current setup will need you to manually add imputation for this country.

Another issue arises if the country you are adding is not as old as the start of the time period of the dataset. To avoid imputing data further back than the start of their existence, you will have to add a line in the id_mat.R script that defines when the nation was created. There are already a number of countries like this in the set, like these examples:

```
id_mat <- subset(id_mat, !(YEAR<1990 & COUNTRY=="Namibia"))
id_mat <- subset(id_mat, !(YEAR<1990 & COUNTRY=="Yemen"))
id_mat <- subset(id_mat, !(YEAR<1993 & COUNTRY=="Eritrea"))
id_mat <- subset(id_mat, !(YEAR<2002 & COUNTRY=="Timor-Leste"))
id_mat <- subset(id_mat, !(YEAR<2011 & COUNTRY=="South Sudan"))
```

Simply add a line like this, replacing the year and the name. The id_mat script is first created with units for all countries for every year in a given period, and here you are merely removing the extra years.

Should a country cease to exist you will also have to add a line for it, but with the limit going the other way.

2.3.3.3 Adding new variables

You can add variables to the mix any way you like, but to preserve order, please do the following:

- Create an R-script in the IDEP_VARIABLES folder.

- Put the data file for the new var in the ORIG_DATA folder, and rename to fit the pattern (i.e. "<varname>_ORIG.<file format>", e.g. "CAKE_CONSUMPTION_ORIG.CSV")

The R-script will automatically be included in the MASTER loop as the MASTER script resources all files in the IDEP_VARIABLES folder. Please do not hard code ANY file paths into the script, such as the location of the data. Use shortened paths from the working directory ("../ORIG_DATA/data.dat", not "C:/.../data/data.dat"). All files necessary to run the script must be put somewhere in the existing system, or given its own subfolder of the main folder.

The script must end with a data frame in the workspace named "data", that should contain ONLY three variables: ISO3C (the three letter ISO country code), YEAR, and the new variable (named whatever you want the new var to be called, in caps).

3. Regression models

The regression model remains largely the same as in previous versions of the GCRI, with the removal of the CON_TREND variable being the only change.

3.1 Design

The setup for the statistical analysis is the same as previous versions, where a best predictor was arrived at by testing many combinations of regression methods and many combinations of variables and interactions. A combination of a logistic model and an OLS model is used. The logistic model is used to calculate predicted probabilities of conflict onset, based on a binary dependent variables derived from the GCRI conflict intensity.

A threshold of 5 is applied to the conflict intensity, with all country-years above this coded as 1, and all below as 0. The predicted probabilities calculate by the model are then used as a basis for deciding which conflicts are to be considered as likely to experience conflict. The threshold is set at 0.3, with all country-years above this predicted to experience conflict.

The OLS model is run directly on the GCRI conflict intensity, using it as the dependent variable. The predicted scores taken from this model are then combined with the probabilities taken from the logistic model. Any country that falls below the probability threshold is scored as 0, while those above are given the predicted score from the OLS model (with any scores over 10 being registered as 10).

Another threshold is then applied to this score, where all countries scoring over 8 being predicted as experiencing highly violent conflict. The same threshold is then applied to the observed conflict intensity. The lists of predicted and observed highly violent conflicts are then compared to evaluate model precision.

3.2 Output

The main purpose of the GCRI is to provide a list of countries predicted to be at risk of experiencing highly violent conflict. The predicted probability is used to filter those countries that are thought to be the most likely to experience a conflict event. The predicted intensities are then used to filter those countries that are thought to have the greatest potential for an occurring conflict to be of a highly violent nature.

Table 8 shows the 20 countries with the highest predicted probability of conflict. These all have a predicted intensities greater than 7.5, meaning they are all expected to not only experience conflict, but also to experience highly violent conflicts.

For an overview of how these countries are placed relative to the rest of the dataset, it is useful to examine the distributions of both the probabilities and intensities.

ISO	COUNTRY	GCRI PROBABILITY	GCRI INTENSITY	GCRI SCORE
IRQ	Iraq	0.99	8.93	8.93
SSD	South Sudan	0.99	10.00	10.00
AFG	Afghanistan	0.99	10.00	10.00
NGA	Nigeria	0.99	9.45	9.45
PAK	Pakistan	0.99	8.42	8.42
SOM	Somalia	0.98	9.46	9.46
SDN	Sudan	0.98	8.94	8.94
MEX	Mexico	0.97	8.62	8.62
IND	India	0.97	7.52	7.52
COD	DRC	0.95	8.54	8.54
LBY	Libya	0.95	7.95	7.95
YEM	Yemen	0.95	9.25	9.25
MLI	Mali	0.94	8.09	8.09
MMR	Myanmar	0.93	8.35	8.35
CAF	CAR	0.93	10.00	10.00
UGA	Uganda	0.93	7.59	7.59
KEN	Kenya	0.92	7.55	7.55
SYR	Syrian	0.92	9.86	9.86
UKR	Ukraine	0.86	9.59	9.59
ETH	Ethiopia	0.81	7.57	7.57

Table 8 - Top 20 countries by predicted conflict probability

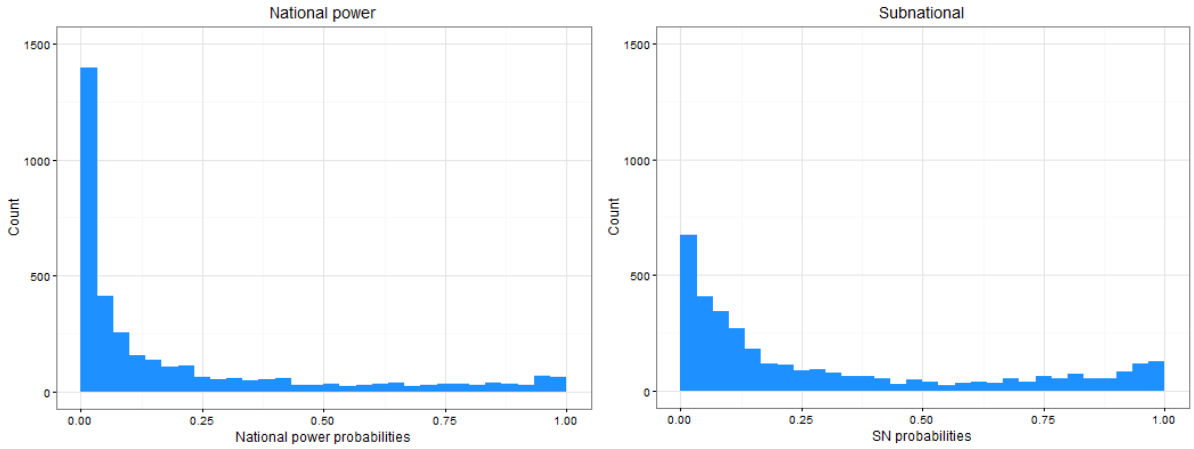


Figure 5 - Distributions of predicted probabilities for both dimensions, for every country-year in the dataset.

Figure 4 shows that the predicted probabilities are both clustered near zero, with Subnational probabilities showing a more flat distribution and greater clustering at the other extreme.

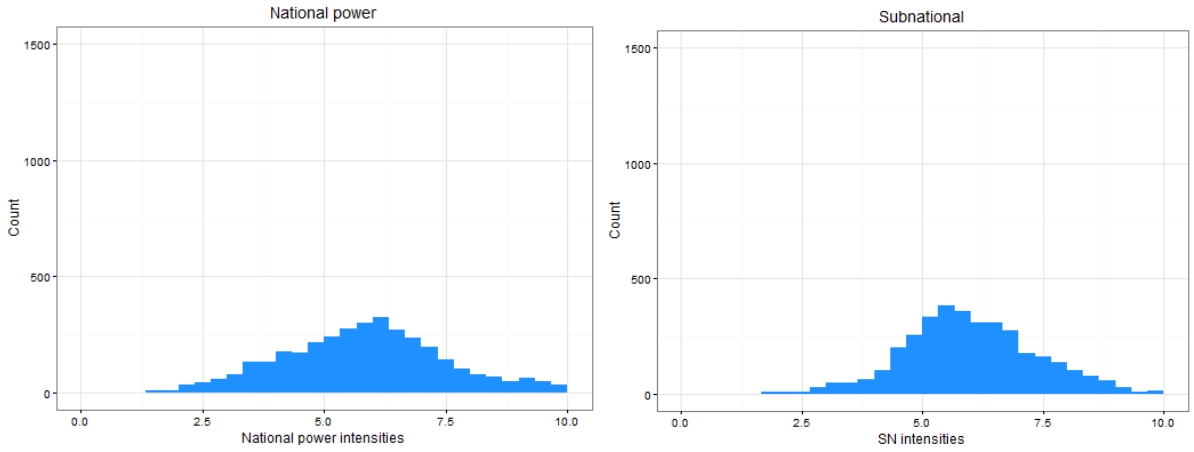


Figure 6 - Predicted conflict intensities for all country-years in the dataset.

The predicted intensities are shown in Figure 5, and are centred on means of 5.8 and 6. They fail to reach the lower part of the scale, both dimensions clustering just over the middle of the scale. This concentration around higher scores is likely a result of the OLS model used to predict them being given only units with scores of >=5 on the dependent variable.

Table 9 shows the coefficients from all the models. Only current conflict intensity (CON_INT) shows an effect that is significant across all four, and the effect is also moderately strong. Inequality (INEQ_SWIID) stands out as the strongest effect in the three first models, and it is also highly significant for these. It does however show a negative sign, as does the GDP per capita variable. These are both contrary to theoretical expectations, even when considering the interactions.

	Model:			
	Logistic		*Linear*	
	(NP)	(SN)	(NP)	(SN)
REG_U	0.201 (-0.409, 0.812)	-0.298 (-0.773, 0.178)	-1.200 (-1.884, -0.516)***	0.268 (-0.216, 0.751)
INEQ_SWIID	-0.952 (-1.728, -0.177)**	-0.897 (-1.393, -0.401)***	-2.161 (-3.318, -1.005)***	0.284 (-0.330, 0.898)
GDP_CAP	-0.401 (-0.958, 0.157)	-0.769 (-1.166, -0.372)***	-1.659 (-2.407, -0.910)***	0.535 (0.086, 0.984)**
REG_P2	-0.034 (-0.105, 0.037)	-0.143 (-0.208, -0.078)***	0.059 (-0.040, 0.158)	0.012 (-0.055, 0.078)
GOV_EFF	0.565 (0.346, 0.784)***	-0.130 (-0.329, 0.068)	-0.191 (-0.478, 0.096)	0.184 (-0.027, 0.394)*
EMPOWER	-0.136 (-0.218, -0.054)***	-0.053 (-0.125, 0.019)	-0.119 (-0.224, -0.015)**	0.070 (-0.002, 0.143)*
REPRESS	0.161 (0.085, 0.238)***	0.177 (0.109, 0.246)***	-0.008 (-0.112, 0.095)	0.011 (-0.064, 0.086)
CON_NB	0.143 (0.103, 0.184)***	0.002 (-0.030, 0.034)	-0.022 (-0.078, 0.035)	-0.025 (-0.061, 0.010)
YRS_HVC	0.090 (0.044, 0.136)***	-0.057 (-0.108, -0.007)**	0.096 (0.042, 0.150)***	0.003 (-0.034, 0.040)
CON_INT	0.220 (0.171, 0.269)***	0.301 (0.253, 0.348)***	0.146 (0.081, 0.211)***	0.158 (0.115, 0.200)***
MORT	0.214 (0.051, 0.378)**	0.323 (0.173, 0.474)***	0.049 (-0.181, 0.279)	0.032 (-0.122, 0.187)
DISPER	-0.043 (-0.081, -0.005)**	0.089 (0.055, 0.122)***	-0.037 (-0.085, 0.011)	-0.057 (-0.094, -0.020)***
HOMIC	-0.043 (-0.116, 0.031)	0.053 (-0.015, 0.120)	0.066 (-0.038, 0.170)	0.113 (0.039, 0.186)***
ETHNIC_NP	-0.011 (-0.120, 0.097)		0.045 (-0.064, 0.155)	
ETHNIC_SN		0.288 (0.228, 0.348)***		0.122 (0.059, 0.184)***
FOOD	-0.042 (-0.155, 0.070)	-0.032 (-0.136, 0.073)	-0.090 (-0.259, 0.079)	-0.107 (-0.211, -0.003)**
POP	0.174 (0.068, 0.280)***	0.316 (0.231, 0.400)***	-0.025 (-0.189, 0.140)	0.362 (0.274, 0.450)***
WATER	0.179 (0.101, 0.257)***	0.035 (-0.034, 0.104)	0.016 (-0.101, 0.132)	-0.023 (-0.099, 0.053)
ECON_ISO	-0.023 (-0.116, 0.070)	-0.026 (-0.109, 0.058)	0.023 (-0.102, 0.148)	-0.065 (-0.155, 0.025)
FUEL_EXP	0.033 (-0.014, 0.080)	0.003 (-0.038, 0.044)	0.087 (0.023, 0.151)***	-0.095 (-0.140, -0.049)***
STRUCT	0.339 (0.215, 0.463)***	0.107 (-0.002, 0.215)*	0.518 (0.334, 0.702)***	0.025 (-0.106, 0.155)
UNEMP	-0.023 (-0.095, 0.049)	-0.078 (-0.136, -0.020)***	-0.159 (-0.270, -0.049)***	0.225 (0.153, 0.297)***
YOUTHBBOTH	0.198 (0.067, 0.328)***	0.059 (-0.045, 0.163)	0.290 (0.092, 0.488)***	-0.227 (-0.346, -0.108)***
CORRUPT	-0.149 (-0.332, 0.035)	0.275 (0.106, 0.445)***	0.275 (0.021, 0.529)**	0.060 (-0.143, 0.263)
REG_U:INEQ_SWIID	0.068 (-0.063, 0.200)	0.076 (-0.031, 0.182)	0.234 (0.080, 0.387)***	-0.091 (-0.203, 0.021)
REG_U:GDP_CAP	-0.075 (-0.168, 0.018)	0.055 (-0.026, 0.137)	0.171 (0.068, 0.274)***	-0.069 (-0.142, 0.003)*
INEQ_SWIID:GDP_CAP	0.091 (-0.030, 0.211)	0.144 (0.060, 0.229)***	0.286 (0.119, 0.454)***	-0.101 (-0.198, -0.003)**
REG_U:INEQ_SWIID:GDP*	-0.0003 (-0.020, 0.020)	-0.016 (-0.034, 0.002)*	-0.034 (-0.057, -0.011)***	0.021 (0.004, 0.038)**
Constant	-7.758 (-11.386, -4.130)***	-3.531 (-5.409, -1.653)***	12.309 (7.264, 17.355)***	1.495 (-1.004, 3.994)
Observations	2,932	2,932	592	891

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9- Regression coefficients and confidence intervals for all regression models.

3.3 Replication manual

The code used for the regression models is mostly the same as used in previous versions, but it has been adapted to the simpler data-handling needed now that the data is supplied complete and not partitioned into separate sets for conflict history, recent data, older data, and so on. The GCRI folder should contain the elements seen in Figure 4.

3.3.1 Folder contents

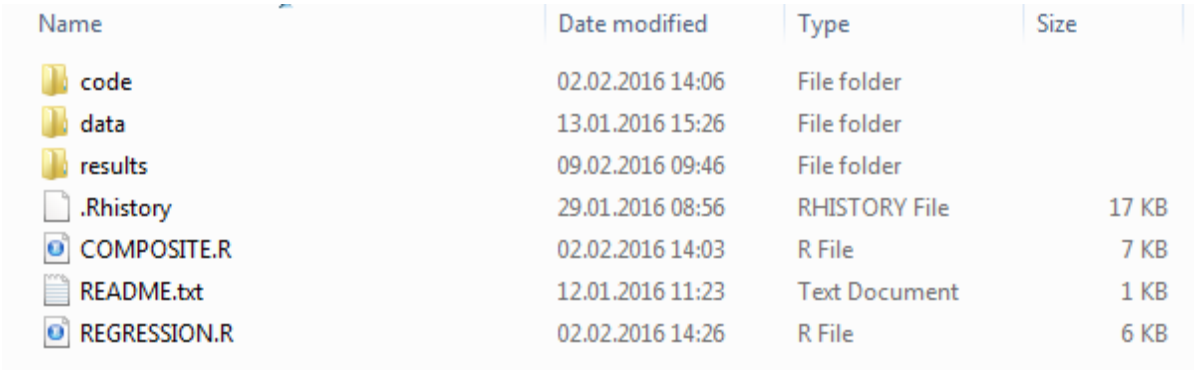


Figure 7 - GCRI folder contents

REGRESSION.R – The R script used to extract the regression predictions, i.e. probabilities, intensities, and scores.

COMPOSITE.R - The R script used to calculate and extract the composite scores.

README.txt – Short instruction text.

code – Folder containing supporting R scripts.

data – Folder containing the GCRI dataset and a dataset with regions.

results – Folder where the results from regression and composite models are stored.

The **code** folder should contain the elements shown in Figure 5.

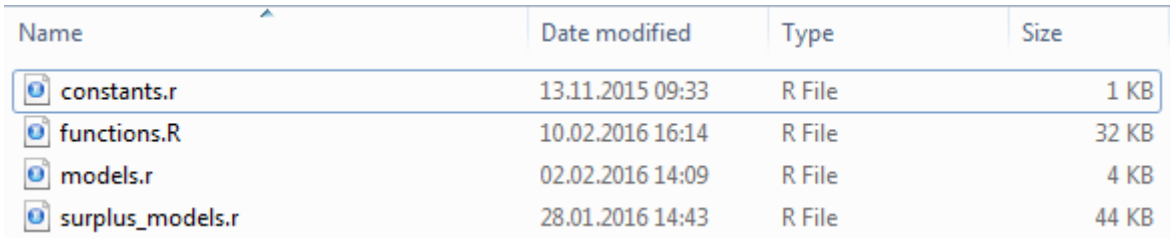


Figure 8 - Code folder contents

constants.r – A short script where you can define thresholds for the various stages of the model.

functions.R – These are functions written by D. Mandrella for the GCRI, with minor modifications. These are needed to perform the regression calculations.

models.r – Contains the single model that is applied to calculate regression results.

surplus_models.r – Some of the other models that were tested.

3.3.2 Walkthrough of the code

To extract the predicted probabilities and intensities from the regression model, you will need to run the REGRESSION.R script.

First you will have to change the working directory to the path where you have the files on your local system.

```
setwd("\\C:\\Users\\Inigo Montoya\\Projects\\GCRI")
```

Once this is done, then all the remaining file paths in the script are set relative to this, and should function if you have kept the folder structure as it was. The next code chunk runs three scripts from the code folder.

```
19 #### Load support functions ####
20
21 source('code/d_functions_revised.r') # Custom functions
22 source('code/models2.r')             # List of models to be run
23 source('code/constants2.r')          # Constants to be used when running model
24
25 #install.packages("pROC") # Run this if you don't have the package installed already
26 library(pROC)     # For Receiver Operating Characteristic curve and Area Under Curve metric
27
28 set.seed(333)
```

functions.r contains a large number of functions that are used to calculate various aspects of the regression results. **models.r** contains the function that runs the regression model, this is where you can find the actual glm and lm models if you want to change variables or interactions. **constants.r** contains the thresholds used by the models, and here you can change the intensity thresholds used to define conflicts, and the probability thresholds for what is predicted as conflicts.

The *pROC* package contains some necessary functions.

The *set.seed()* command makes R start drawing random numbers from a fixed point. This is to ensure that the results can be replicated exactly.

The next step is to load the dataset, and to prep the conflict variables that depend on the threshold set in **constants.r**.

```
30 #### Load and prep data ####
31 load("data/gcri.data.Rdata")
32
33 gcri.data         <- gcri.data[with(gcri.data, order(ISO, -YEAR)), ]
34
35 # Compute onsets using the thresholds sourced from the constants script
36 gcri.data$VC_Y_NP  <- ifelse(gcri.data$Intensity_Y_NP  >= kViolentConflictIntensityThreshold, 1, 0)
37 gcri.data$VC_Y_SN  <- ifelse(gcri.data$Intensity_Y_SN  >= kViolentConflictIntensityThreshold, 1, 0)
38 gcri.data$VC_Y4_NP <- ifelse(gcri.data$Intensity_Y4_NP >= kViolentConflictIntensityThreshold, 1, 0)
39 gcri.data$VC_Y4_SN <- ifelse(gcri.data$Intensity_Y4_SN >= kViolentConflictIntensityThreshold, 1, 0)
40 gcri.data$HVC_Y_NP <- ifelse(gcri.data$Intensity_Y_NP  >= kHighlyViolentConflictIntensityThreshold, 1, 0)
41 gcri.data$HVC_Y_SN <- ifelse(gcri.data$Intensity_Y_SN  >= kHighlyViolentConflictIntensityThreshold, 1, 0)
42
43 # Add regions
44 gcri.data <- addRegions(gcri.data)
```

Line 31 simply loads the dataset that is found in the data folder.

Line 33 orders the data by ISO code, and then ascending by observation year.

Lines 36 to 41 then creates dummy variables for Violent Conflict (VC) in the observation year, VC in the next four years after the observation year, and for Highly Violent Conflict (HVC) in the observation year.

Line 44 adds regions to the data using the dataset in the data folder.

The data is now ready, and the models can be run.

```
46  #=================================================================================================
47  #### Evaluation of model
48  #=================================================================================================
49  setwd("results")
50
51  models <- list()
52  models[['REG_UandINEQ_SWIIDandGDP_CAPNewVarsBoth']]   <- REG_UandINEQ_SWIIDandGDP_CAPNewVarsBoth
53
54  # Subset units with data on dependent variable
55  train <- subset(gcri.data, YEAR<=2010)
56
57  nv.cv.model.results <- runModels(list.of.models = models, dat = train, cross.validation = TRUE)
58  compareModels(nv.cv.model.results, save.as.csv = TRUE, custom.name = 'new-vars__with-cv')
```

First, the working directory is changed so that the output is saved in the results folder. A list is then created, and the function that contains the glm and lm models is added to it. If you want to test more than one model, for example those in surplus_models.r, you can add additional models here after sourcing the surplus_models script or creating your own.

The dataset is then subset to only the units that have data on the dummy variable for conflict 4 years ahead in time.

The *runModels()* function on line 57 is used to evaluate the model by cross-validation. It creates an object that is then saved to a .csv by the *compareModels()* function. This creates two .csv files like these:

They contain some evaluation metrics that give the predictive performance of the model. The next step is to extract the actual predictions.

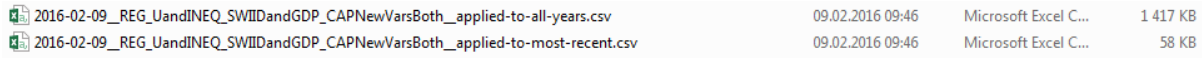
```
#=================================================================
#### Extract GCRI scores for all countries in dataset
#=================================================================

# Subset final year of the set
most.recent.data <- subset(gcri.data, YEAR==max(gcri.data$YEAR))

# Extract scores for final year
applied.models.static <- applyModels(list.of.models = models,
                                     train.data = train,
                                     apply.data = most.recent.data)
saveAppliedModelsOutput(applied.models.static)

# Extract scores for all years in set
applied.static.all.years <- applyModels(list.of.models = models,
                                        train.data = train,
                                        apply.data = gcri.data)
saveAppliedModelsOutput(applied.static.all.years, applied.to = 'all')
```

The last year is subset, and then the model is trained on the full set and applied to this data using the applyModels() function. The model can also be applied to the whole period to see development over time. Doing both should result in this output in the results folder:



Finally, the coefficients of the four regressions can be extracted, either as text printouts in the R console, or as html/LaTeX.

```r
###### Extract coefficients ########

np_model <- REG_UandINEQ_SWIIDandGDP_CAPNewVarsBoth(train, threshold = "static", conflict.dimension = "np", compute.metrics=F)
sn_model <- REG_UandINEQ_SWIIDandGDP_CAPNewVarsBoth(train, threshold = "static", conflict.dimension = "sn", compute.metrics=F)

np_model$logit.model$coefficients
sn_model$logit.model$coefficients

np_model$linear.model$coefficients
sn_model$linear.model$coefficients

# html
library(stargazer)
stargazer(np_model[["logit.model"]],sn_model[["logit.model"]],np_model[["linear.model"]],sn_model[["linear.model"]],
          type="html", single.row = T, report = "vcs*",ci=T)

stargazer(np_model[["logit.model"]],sn_model[["logit.model"]],np_model[["linear.model"]],sn_model[["linear.model"]],
          type="html", single.row = T, report = "vc",ci=T)
```

4. Composite model

To provide a more intuitive alternative to the regression model, a composite indicator has been developed. The composite indicator was to use the same theoretical framework as the regression model, with the same variables in the same groups.

4.1 Model selection

Several alternatives methods of weighting and grouping the variables were explored:

- Weights derived from PCA
- Weights derived from regression coefficients
- Arithmetic mean
- Geometric mean
- All variables together, in groups, different combinations of weights on different levels

A PCA was first done to ascertain whether the variables were fit to be combined in such a manner. Only two variables were found to be problematic, with factor loadings in the very low negatives compared with the rest. Testing was carried out both with these two variables included, and without them. Keeping the variables had no clear adverse effect on the predictive power on the final averaging model chosen (The variables were kept to preserve the political narrative).

First, the factor loadings from the PCA were used to take a weighted average of all the variables. This was also done without the conflicting variables, and without variables that had low factor loadings. The following score was then divided by 10, and treated as a probability. The same was done using arithmetic and geometric means. The methods were also combined, dividing the data first into smaller groups along the risk areas and components. Lastly, the regression coefficients were also used to derive weights for a weighted average, and combinations of the coefficients and their standard errors.

The "probabilities" were then tested up against observed conflict data, and evaluated by their ROC AUC, PR AUC, and precision level at certain recall levels. The regression based weights performed by far the worst, with only 2/3 the precision of the others. The remaining models very mostly clustered at the same precision levels, with only +/- 0.02 differences in the various measures. A few, the simpler averaging models and one of the geometric averaging models, were consistently slightly above the others. The PCA models were discarded as the weights in some cases discarded variables entirely, it required substantially more work, and did not perform better than simpler models.

Table 7 illustrates how a simple average performed almost on par with the more complex averaging combinations. The geometric mean of all variables was noticeably worse than the models containing simple averages, while the remaining combinations were mostly similar. The models with a simple average in the last step did perform somewhat better than those with a geometric average, suggesting that single extreme values should not be emphasized. Of the two best performing models, the simple grouped mean method was chosen, as its calculation is much simpler and intuitively understood.

4.2 Model results

The method reduces the number of variables in steps, from 24 to 10, from 10 to 5, and from 5 to 1, using the arithmetic mean of the theoretically defined groups (See example on next page). Due to the smoothing effect of averaging in this manner, the composite scores for the two conflict dimensions are almost identical.

Method	TPR	PPV	PRAUC	PREC75	PREC80	PREC85	
Regression	*0.74*	*0.66*					
Average	0.762	0.496	0.672	0.51	0.47	0.43	NP
Geometric r	0.943	0.324	0.659	0.49	0.436	0.396	
Grouped m	0.662	0.577	0.694	0.512	0.475	0.444	
Averaging g	0.726	0.534	0.689	0.509	0.481	0.443	
Geom of gr	0.787	0.485	0.69	0.511	0.481	0.439	
Geom of gr	0.723	0.523	0.693	0.507	0.47	0.436	
Regression	*0.77*	*0.71*					
Average	0.677	0.564	0.698	0.516	0.481	0.458	SN
Geometric r	0.865	0.421	0.679	0.489	0.465	0.433	
Grouped m	0.587	0.669	0.697	0.513	0.498	0.454	
Averaging g	0.635	0.621	0.703	0.528	0.491	0.45	
Geom of gr	0.682	0.564	0.696	0.508	0.484	0.44	
Geom of group avg	0.646	0.603	0.691	0.505	0.478	0.438	

Table 10 - Results of cross validation. The first two columns are the results from using the same metrics calculation function as was used with regression results. This uses a fixed threshold of 7, and units with composite scores over 5 are judged to be conflicts. The remaining four columns were calculated in-sample. The Precision Recall AUC gives overall performance over all probability thresholds. The last three give the precision level at recall levels of .75, .80, and .85, with the probability threshold being fluid.

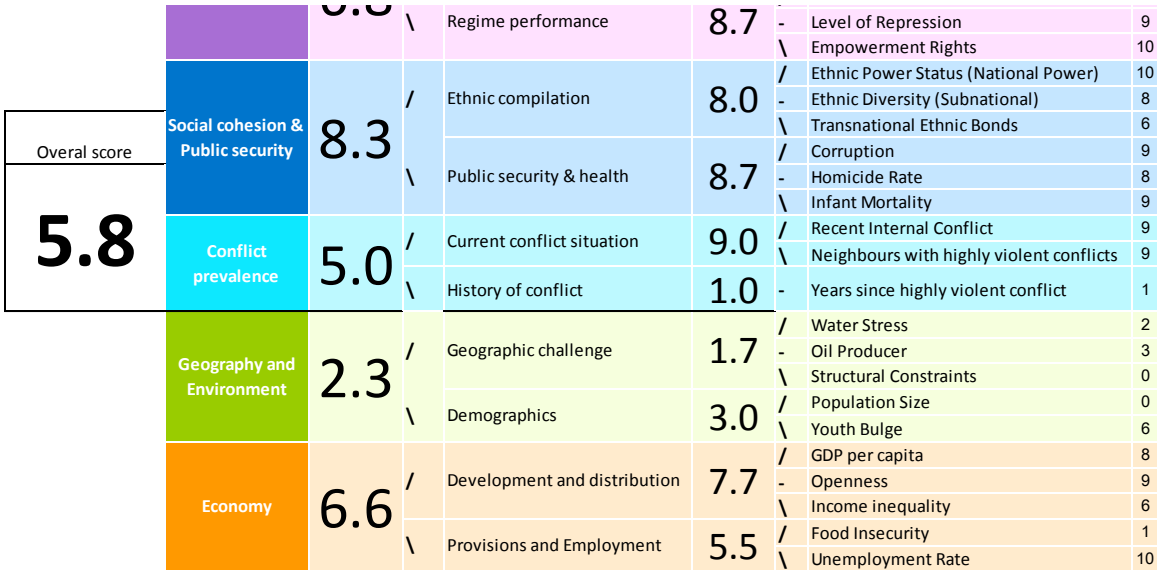


Figure 9 - Composite Index grouping of variables. Each step averages the groups components, until a single score is left.

The TPR and PPV metrics above were calculated using the same fixed cutoff point for conflict onset as applied to the regression results. While they clearly show that all models are worse than regression, the tradeoff between the two metrics varies greatly. Further tests were run, with the constant being changed from probability cutoff to TPR levels close to that achieved by the regression model. The columns marked PREC75, PREC80, and PREC85 show the precision level of the model when the TPR is at .75, .80 and .85. The Area Under the Precision Recall Curve (PRAUC), is also reported.

Figure 7 shows how the two composite scores correlate with their respective regression probabilities, intensities, and scores. It is clear that the regressions and the composite model creates very different results, but also that they are in some agreement. For the probabilities, the correlation is clearer for NP than for SN. The intensities reveals a potential strength in the composite model, as it gives the ANZAC and EFTA countries at very low scores where the NP regression model gives medium/high intensity predictions.

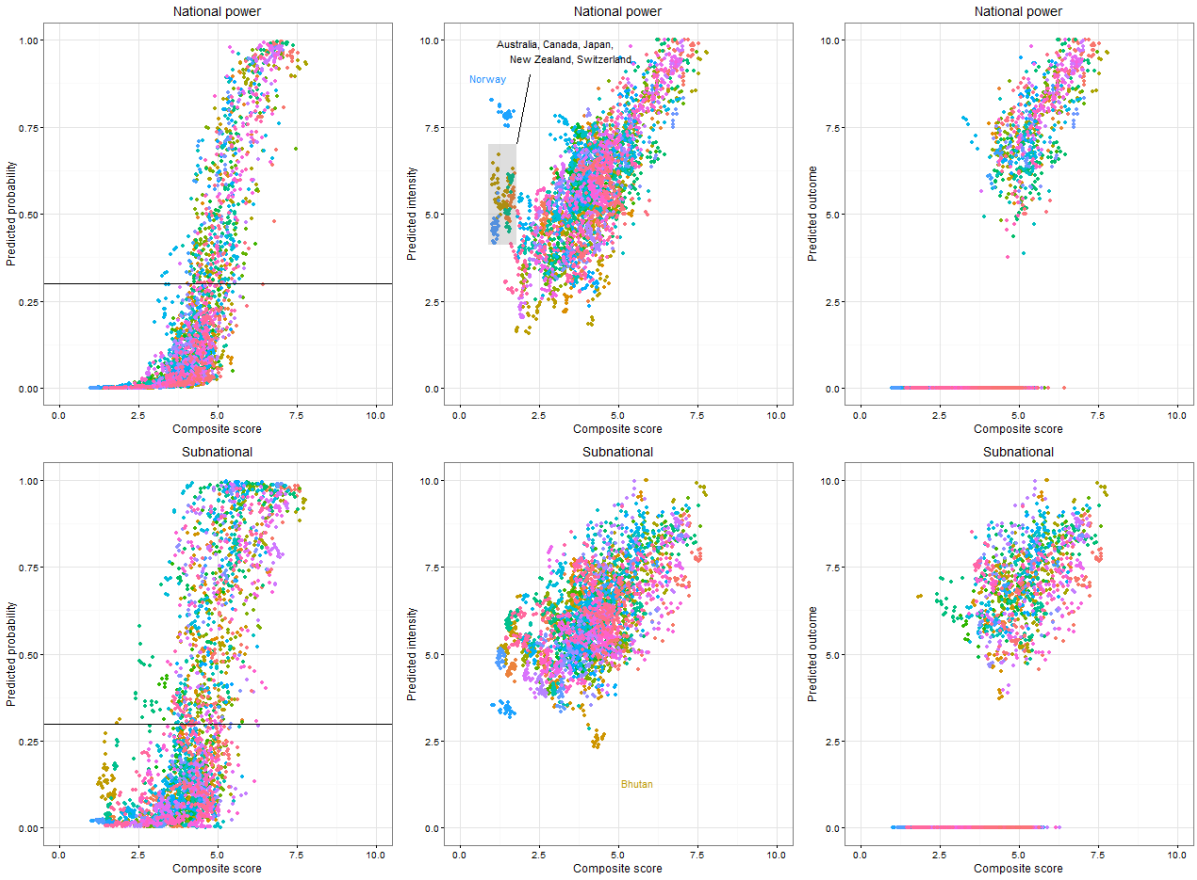


Figure 10 - Composite vs Regression. The two plots on the left show the composite scores versus the predictied probabilities of the logistic regression model. The middle plots show the composite score versus the predicted intensities from the OLS model. The plots on the left show the composite score versus the GCRI score (i.e. the intensities of the units with a probability over 30%).

4.3 Replication manual

This section walks the reader through the code that replicates the composite score part of the GCRI. For the files that should be present, please see section 3.3.1, as the composite and regression come in one package. The COMPOSITE.R script is the only one you will have to open.

4.3.1 Walkthrough

Open **COMPOSITE.R**, and input your own working directory in the *setwd()* command in the beginning of the script.

```
setwd("\\C:\\Users\\Billy Pilgrim\\Projects\\GCRI")
```

The data is then loaded as in the **REGRESSION.R** script, before the main loop starts.

Rather than write the whole script twice, the script is looped over twice, once for each conflict dimension. This is done because the two dimensions use different ethnic variables. First, a vector containing he name of the dimensions is created. Second, the loop is started, going through the values of the dimensions list.

```
25  # Make a vector containing the conflict dimensions to loop over
26  dimensions <- c("sn","np")
27
28  # Loop over conflict dimensions
29  for (condim in dimensions){
```

The loop starts by subsetting the data based on the dimension.

```
31    # Subset the correct ethnic variable depending on dimension
32    if(condim=="np"){
33      mat_test <-
34        subset(
35          gcri.data, select = c(
36            REG_U, REG_P2,GOV_EFF,EMPOWER,REPRESS,
37            CON_NB,CON_INT,YRS_HVC,
38            MORT,DISPER,HOMIC,ETHNIC_NP,CORRUPT,
39            INEQ_SWIID, GDP_CAP, ECON_ISO, FOOD, UNEMP,
40            POP,STRUCT,YOUTHBBOTH, WATER,FUEL_EXP
41          )
42        )
43    }else if (condim=="sn"){
44      mat_test <-
45        subset(
46          gcri.data, select = c(
47            REG_U, REG_P2,GOV_EFF,EMPOWER,REPRESS,
48            CON_NB,CON_INT,YRS_HVC,
49            MORT,DISPER,HOMIC,ETHNIC_SN,CORRUPT,
50            INEQ_SWIID, GDP_CAP, ECON_ISO, FOOD, UNEMP,
51            POP,STRUCT,YOUTHBBOTH, WATER,FUEL_EXP
52          )
53        )
54    }
```

Only the relevant data variables are taken, with the only difference between dimensions being the ETHNIC_NP/SN variable. After this, the datasets are split into frames containing only the variables of each concept.

```r
# split risk areas into concepts
pol_mat_test1  <-
  as.matrix(subset(mat_test, select = c(REG_U,REG_P2)))
pol_mat_test2  <-
  as.matrix(subset(mat_test, select = c(GOV_EFF,EMPOWER,REPRESS)))

sec_mat_test1  <-
  as.matrix(subset(mat_test, select = c(CON_NB,CON_INT)))
sec_mat_test2  <-
  as.matrix(subset(mat_test, select = c(YRS_HVC)))

# Subset correct ethnic variable depending on dimension
if(condim=="np"){
  soc_mat_test1  <-
    as.matrix(subset(mat_test, select = c(DISPER,ETHNIC_NP,CORRUPT)))
}else if (condim=="sn"){
  soc_mat_test1  <-
    as.matrix(subset(mat_test, select = c(DISPER,ETHNIC_SN,CORRUPT)))
}
```

The loop again distinguishes between dimensions when creating the set for the first social concept.

After grouping the variables in concepts, the means of the variables in them are taken and the two means are bound together to make a matrix for each risk area.

```r
# Bind the two concepts of each risk area into a frame
pol_mat_test <- cbind(
  rowMeans(pol_mat_test1),
  rowMeans(pol_mat_test2)
)
sec_mat_test <- cbind(
  rowMeans(sec_mat_test1),
  rowMeans(sec_mat_test2)
)
soc_mat_test <- cbind(
  rowMeans(soc_mat_test1),
  rowMeans(soc_mat_test2)
)
econ_mat_test <- cbind(
  rowMeans(econ_mat_test1),
  rowMeans(econ_mat_test2)
)
geo_mat_test <- cbind(
  rowMeans(geo_mat_test1),
  rowMeans(geo_mat_test2)
)
```

The average of the two concepts of each area is then taken, and the resulting 5 risk area scores are bound together. The average of these 5 scores are then the final scores, with each row representing the country year in question.

```r
# Average the two concepts in each area to create a single score for each risk area
group_scores_averaged <- cbind(
  rowMeans(pol_mat_test),
  rowMeans(sec_mat_test),
  rowMeans(soc_mat_test),
  rowMeans(econ_mat_test),
  rowMeans(geo_mat_test)
)

# Average the risk areas into a single score for each country-year
mean_mean_scores    <- as.data.frame(rowMeans(group_scores_averaged))
```

The problem now is that we only have a long list of numbers. Fortunately, everything is still in the same order as when we started, so we simply bind the scores to the ID variables from the dataset we started with.

```
123   # Bind ID variables to the scores
124   mean_mean_scores    <- as.data.frame(cbind(
125     as.character(gcri.data$ISO),
126     as.character(gcri.data$COUNTRY),
127     gcri.data$YEAR,mean_mean_scores))
128   colnames(mean_mean_scores) <- c("ISO","COUNTRY","YEAR","SCORES")
129   rownames(mean_mean_scores) <- paste(mean_mean_scores$ISO, mean_mean_scores$YEAR, sep="")
```

The loop then stores the resulting frame, creating one for each dimension. These are then combined into one frame, the variables are named, and the results are stored using the write.csv() function. The .csv files should appear in the results folder under the name given in the write.csv() function, set to "composite_scores_mostrecent" or "composite_scores_allyears" by default.

NOTE: The clumsy structure of the code is a remnant from the test stage when different methods were used between levels, requiring output at both concept and risk area levels, not only the final score. The rowMeans() function can easily be exchanged with other functions at some levels, or for some specific concepts or areas. Alternatives could be rowMaxs (MatrixStats package) for the highest scoring variable

5. Conclusion

The latest version of the GCRI has made great progress towards making the GCRI more open, both for the purposes of scientific transparency and reproducibility, and in terms of making the output more accessible.

While there is still much room for improvement, the changes in data management have given more reliable data. The use of R scripts limits the possibility for human error, or at least makes it easier to discover and correct such errors as all actions are recorded.

The inclusion of the composite indicator in the GCRI adds a tool that gives a more easily interpretable presentation of the raw data, which also appears to give more realistic estimates of conflict potential in countries that have a low conflict probability. The tests of composite indicators also bring a valuable lesson, in that they indicate that high values on single indicators are not important in determining conflict risk. The models that included the "smoothing" effect of an arithmetic average all outperformed the models that relied solely on geometric averaging.

Further work should focus on improving the imputation methods through interpolation and extrapolation where relevant. Having reliable data is crucial in uncovering actual causal mechanisms and not random patterns in noisy data. In connection with this, effort should also be put into finding improved data for the variables with the most serious availability problems. This not only adds more "real" information, but also improves the accuracy of imputed data.

The possibility of having separate models for each dimension should also be explored. This would maximize the predictive power of each, and should not be to difficult to implement by modifying the existing framework.

References

Collier, Paul, and Anke Hoeffler. "Greed and grievance in civil war." *Oxford economic papers* 56.4 (2004): 563-595.

Collier, Paul, Anke Hoeffler, and Dominic Rohner. "Beyond greed and grievance: feasibility and civil war." *Oxford Economic papers* (2008): gpn029.

Eck, Kristine and Lisa Hultman. 2007. "Violence Against Civilians in War." Journal of Peace Research 44(2)

Fearon, James D., and David D. Laitin. "Ethnicity, insurgency, and civil war." *American political science review* 97.01 (2003): 75-90.

Fox, Jonathan. "The rise of religious nationalism and conflict: Ethnic conflict and revolutionary wars, 1945-2001." *Journal of peace Research* 41.6 (2004): 715-731.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002) Armed Conflict 1946-2001: A New Dataset. *Journal of Peace Research* 39(5).

Gurr, T. R., Harff, B., Levy, M., Dabelko, G. D., Surko, P. T., & Unger, A. N. (1999). State failure task force report: Phase II findings. *Environmental Change & Security Project Report*, (5), 50.

Hegre, Håvard. "Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992." *American Political Science Association*. Vol. 95. No. 01. Cambridge University Press, 2001.

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., & Urdal, H. (2013). Predicting Armed Conflict, 2010–20501. *International Studies Quarterly*, *57*(2), 250-270.

Hegre, Håvard, and Nicholas Sambanis. "Sensitivity analysis of empirical results on civil war onset." Journal of conflict resolution 50.4 (2006): 508-535.

Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010). "The Worldwide Governance Indicators: Methodology and Analytical Issues". World Bank Policy Research Working Paper No. 5430

Regan, Patrick M., and Daniel Norton. "Greed, grievance, and mobilization in civil wars." *Journal of Conflict Resolution* 49.3 (2005): 319-336.

Sundberg, Ralph, Kristine Eck and Joakim Kreutz, 2012, "Introducing the UCDP Non-State Conflict Dataset", Journal of Peace Research, March 2012, 49:351-362

UCDP Battle-Related Deaths Dataset v.5-2015, Uppsala Conflict Data Program, www.ucdp.uu.se, Uppsala University

List of abbreviations and definitions

BRD – Battle related deaths
GCRI – Global Conflict Risk Index
NSC – Non-state conflict
OSV – One-sided violence
UCDP/PRIO – Uppsala Conflict Data Program/Peace Research Institute Oslo

List of figures

List of tables

How to obtain EU publications

Our publications are available from EU Bookshop (http://bookshop.europa.eu),
where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents.
You can obtain their contact details by sending a fax to (352) 29 29-42758.

LB-NA-27908-EN-N

JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

doi:10.2788/705817

ISBN 978-92-79-58227-1