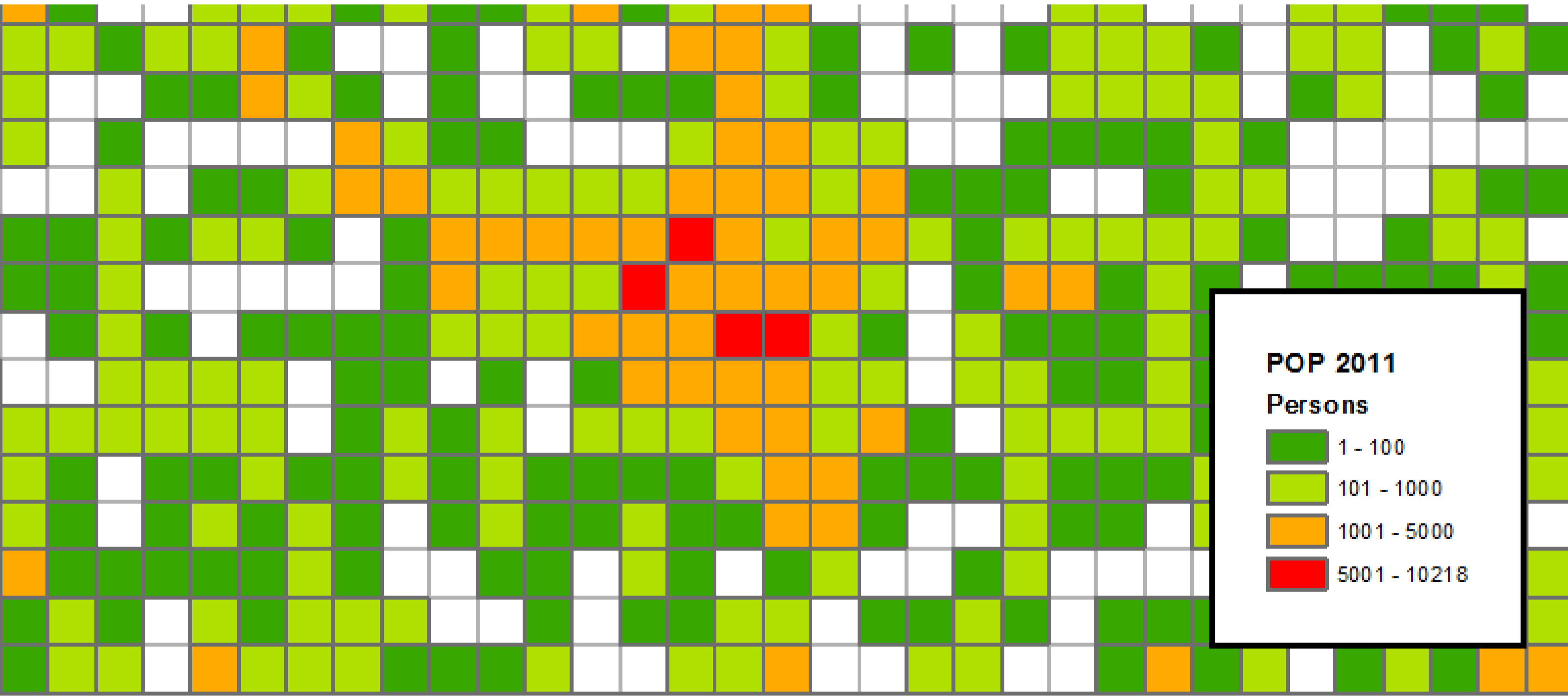
European
Commission

JRC TECHNICAL REPORT



Urban and Regional Built-up Analysis (URBA)

Production of a population grid in Europe (1 km of resolution)

Sergio FREIRE, Stamatia HALKIA, Martino PESARESI, Daniele EHRLICH

2015

Report EUR 27482

Joint Research Centre

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen (IPSC)
Global Security and Crisis Management Unit

Contact information
Sergio FREIRE
Address: Joint Research Centre, Via E. Fermi 2749, I-21027 Ispra (VA) Italy, TP 267

E-mail: sergio.freire@ec.europa.eu
Tel. +39 0332 785503

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

JRC97911

EUR 27482

ISBN 978-92-79-52068-6 (PDF)
ISBN 978-92-79-52067-9 (print)

ISSN 1831-9424 (online)
ISSN 1018-5593 (print)

doi:10.2788/68127

Luxembourg: Publications Office of the European Union, 2015

Printed in Ispra *(Italy)*

Abstract

Mapping settlements in Europe is an on-going research project performed in collaboration with the Directorate-General for Regional and Urban Policy (DG REGIO). The Joint Research Centre (JRC) expertise in automatic extraction of information from satellite imagery for the mapping of human settlements has been further enhanced with state-of-the-art population disaggregation methodologies with an aim to provide fine resolution population data across Europe, where detailed population data do not exist, to enable evaluation of current, and planning of future European policies in the context of territorial cohesion. The work described in this paper, summarizes the background, methodology, experiments, reference data, and results, which has been performed towards a wall-to-wall population grid in Europe at 1 sq.km of resolution.

Table of Contents

List of Tables

List of Figures

Executive Summary

The research work was performed in the context of an administrative arrangement between the European Commission services: the Directorate General for Regional and Urban Policy (DG REGIO), and the Joint Research Centre (JRC), JRC-Nr.33284-2013, during April 2014-March 2015. The report summarizes the background, methodology, experiments, reference data, and research results, related to the production of a population grid at 1 km of resolution.

Research and development, included bibliographical review of state-of-art population disaggregation methodologies, extensive experimental testing using case studies at regional and national level, testing modelling methods, cost-benefit analyses on methodological approaches, review of input data from various sources, and finally production of population grids for three European countries, Cyprus, Luxembourg and Malta, as requested by the customer (DG REGIO)[1]. Land use data from the European Urban Atlas and CORINE LAND USE LAND COVER (CLC06rv2) were used as ancillary reference to weight population according to the functional characteristics of the settlement area.

Particular attention was given to the masking or weighting of certain land use classes according to the available data provided for model validation. A significant part of the research involved was dedicated to statistical evaluation of alternative methods of weighting and masking land use classes.

Preservation of population counts was observed, according to the requirements. Issues regarding the quality and types of population input data, as well as discrepancies in the spatial domain, were challenging. In the case of Cyprus, particular attention was given to the specific political context. In the case of Luxembourg, the availability of country-wide reference data (Urban Atlas), yielded optimized results.

Overall the success of this work has relied on the quality of the input data provided and the availability of ancillary land use data in Europe.

[1] Eight European countries, in total, by May 2015.

Introduction

The work described in this report concerns the methodological development and production of a population grid at 1 km^2 of resolution for Europe for the year 2011. This work aims at improving the following task: Update and improvement of a population grid (2011) using a disaggregation methodology. Specifically, WP 2.2.1 concerns the production of a population grid (2011) at 1 km^2 of resolution for Europe, using census tract data, 1 km^2 grid or data at the level of municipalities (LAU2), whatever best is available in each country.

It is expected that these results can add significant value to European Commission cohesion policies in territorial and regional planning, as well as provide disaggregated population data for Eurostat, where census tract or administrative-level data is not available.

Additionally, the population grid will be used as input to the GHSL settlement model for Europe. The settlement model currently defines settlement types in Europe, based on the OECD city definition, and additional definitions provided by the GHSL model [1], [2]. The GHSL definitions for rural and peri-urban settlements will refine the sparsely inhabited areas used by the European Commission to describe rural settlements, by the physical characteristics of the settlement and population density.

This report is structured in the following way: In section 1 the context is presented and the task and the goals are framed; section 2 provides some background and state-of-the-art information on population disaggregation; section 3 reports on some of the preparatory and exploratory work conducted in order to meet the goals; section 4 presents the selected methodology and developed tools; section 5 shows some of the obtained results, while section 6 offers conclusions and future outlook.

DG REGIO requirements

In the context of an administrative arrangement between the Joint Research Centre (JRC) and the Directorate General for Regional Policy (DG REGIO) on the built-up detection of human settlements in Europe, JRC was requested to contribute to an updated and improved population grid in Europe.

Specifically, workpackage 2.2.1 targeted the production of a population grid (2011) at 1 km^2 of resolution for Europe, using census tract data, 1 km^2 grid or data at the level of municipalities (Local Area Unit - LAU2), whatever best is available in each country. Therefore the adopted approach should be sufficiently flexible to accommodate input population totals referenced to different spatial geometries. Additionally, the task was to be carried out within some bounding conditions, namely:

a) Aim at a residence-based population representation
b) Preserve population input totals (volume preservation)
c) Take advantage of GHSL's detailed mapping of built-up areas

Table 1 shows the list of countries selected for GHSL-based 1 km^2 population grid production, as of January 2015.

			GRID GEOSTAT1B 2011 V1_0			selected for GHSL-based 1 km grid		
CNTR_ID		CLC06_r_v2	Aggregated	Hybrid	Disaggregated AIT	Initially	Oct. 2014	Jan. 2015
EL	Ellada				X		Y	?
HR	Hrvatska	Y			X	Y		
IT	Italia	Y			X	Y	Y	?
CY	Kýpros	Y			X	Y	Y	Y
LV	Latvija	Y			X	Y	Y	
LT	Lietuva	Y	X			Y	Y	
LU	Luxembourg	Y1			X	Y	Y	Y
MT	Malta	Y			X	Y	Y	Y
RO	România	Y			X	Y		

Table 1. Countries selected for GHSL-based 1 km^2 population grid (in yellow) and available datasets

Background on population disaggregation

Population censuses normally provide accurate information on the characteristics and number of residents for administrative or finer enumeration areas. However, these data are not uniformly and readily available across European Union member states. Additionally, these data sets are typically available as a total count for units varying widely in size and shape, while frequently residents occupy only specific zones of these units, at different densities [3]. Making these data available as regular grids facilitates modelling and spatial analysis, and mitigates biases such as the modifiable areal unit problem (MAUP) and the related ecological fallacy issue. Furthermore if the grids are sufficiently fine, re-aggregation to different geometries (e.g. natural hazard zones) is greatly facilitated and expedited.

The goal of population disaggregation is to improve the representation of population distribution and densities, by reallocating (distributing) existing population counts from census or administrative zones to areas where it is more likely to be present, increasing the spatial discrimination and/or the resolution (detail) of the distribution. This process typically relies on geographical modeling to disaggregate and refine population distributions.

The use of ancillary data that is related to population distribution, such as land use (especially if compatible in time, scale, and semantics) can be valuable to further inform the disaggregation process by allowing the discrimination of distinct functional areas. One of the most effective ways to take advantage of ancillary data is through dasymetric mapping and areal interpolation approaches. Dasymetric mapping is a cartographic technique, originally developed for population mapping [4], which aims at limiting the distribution of a variable to the areas where it is present, by using related ancillary information in the process of areal interpolation [5]. Dasymetric mapping approaches typically require the following geographic information:

- Zones with population counts (i.e., source zones)
- Ancillary information – proxies or covariate spatial layers that correlate well with presence of population (e.g. Land Use, Built-up, etc.)

Many methods have been proposed for population disaggregation – for reviews of existing approaches, refer to [6] and [7]. Among the recent proposals aimed at raster-based disaggregation, is the Intelligent Dasymetric Mapping (IDM) method developed by Mennis 2003 [8], and Mennis and Hultgren 2006 [9]. This is basically a multiclass dasymetric mapping technique using empirical sampling to derive reference population densities for ancillary categorical classes (e.g., LULC) to inform the disaggregation process. In tests, it has outperformed areal weighting and binary dasymetric mapping. Langford [7], testing small area population estimation techniques, has concluded that the strongest performance was obtained by combining dasymetric mapping with data on the locations of buildings.

Recent approaches to population disaggregation in Europe include those implemented by [10] and [11], using European-wide geographic information and dasymetric approaches. Steinnocher et al. [10] have disaggregated census population to the European Soil Sealing Layer (ESSL), after masking this raster layer with a combination of selected CORINE Land Cover (CLC) classes and transportation data. Results showed underestimation in urban centers and rural areas. Batista e Silva et al. [11] have produced a 100-m resolution population layer for 2006 by disaggregating census data to a selected set of CLC refined classes ('Urban fabric') by using weights derived from the average soil sealing degree per LULC class.

Notwithstanding some methodological innovations recently proposed, empirical evidence suggests that the greatest determinant of disaggregation accuracy is the size of the source zones, followed by the quality of the covariate. A simple regression analysis of source zone median size (communes) and results of best method (based on Relative Total Absolute Error - RTAE) for seven countries, as reported in [11] seems to confirm the importance of this factor (Figure 1). Correlation coefficient (*r*) is 0.74.

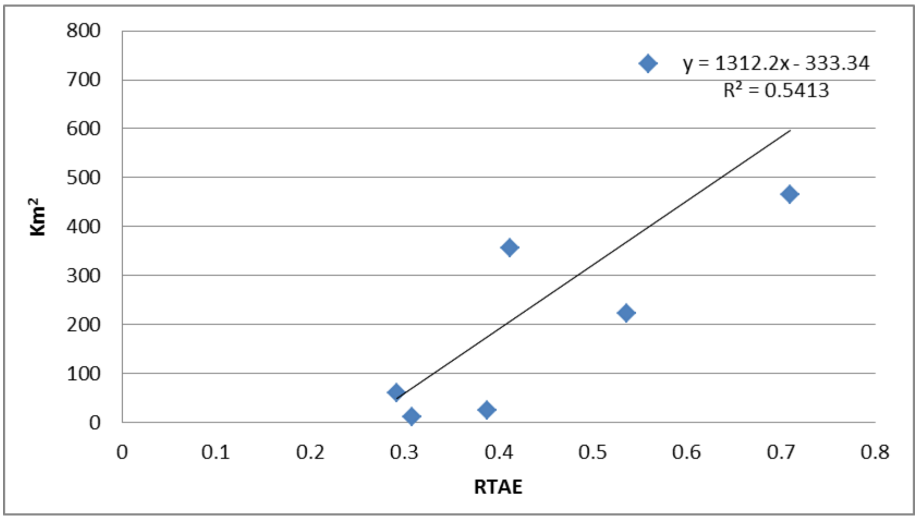


Figure 1. Correlation between source zone size and disaggregation accuracy for results reported in [11]

Methodological development and experimentation

In this WP, significant preparatory and exploratory work was conducted. This work aimed at both exploring the data sets involved and testing different methodological approaches. The experiments were conducted at different scales and resolutions, initially limited by the available data and progressively expanding their geographical scope as more data became available. This stage allowed to familiarizing oneself with strengths and weaknesses of the data sets and possible methodologies applicable.

Three types of data sets were available and were involved for population disaggregation in Europe:

- Census resident population (source zones), by enumeration or administrative area
- GHSL Built-Up layer (target zones)
- Land Use / Land Cover (LULC) maps (ancillary data)

These data sets can have quite different characteristics, as illustrated in Table 2 (Census from Portugal):

DATA SET	Name	Date	Structure	Data type	Cell size	MMU (ha)
Population	Census blocks	2011	Vector	Numerical (counts)	--	0.016 (PT)
Population	LAU2	2011	Vector	Numerical (counts)	--	5.2 (PT)
Built-up	GHSL BU	2011	Raster	Numerical (Ratio)	10 m	--
Land use	CLC06_r_v2	2006	Raster	Categorical	100 m	1 – 25

Table 2. Types of input data sets involved and their characteristics

The GHSL Built-Up layer is a seamless 10m-resolution raster mosaic representing, for each cell, the ratio of its surface area covered by building structures [2] (see Fig. 2).

Figure 2.GHSL Built-Up layer (workflow 7)

A European-wide 100-m resolution LULC layer, named CLC06_r_v2, was produced by Batista e Silva et al. [12] by combining CLC2006 with additional geoinformation to increase thematic and spatial detail (Figure 3 and Table 3). This refined CLC has 45 classes, a minimum mapping unit of 1 ha for Artificial surfaces and Inland waters classes and 25 ha for the remaining. However, while GHSL mostly represents 2011 and 2012, CLC06_r maps LULC in 2006.

Figure 3. CLC Refined 2006 version 2 layer (CLC06_r_v2.tif)

CLC_r Code	CLC Code	Class name
0	0	NODATA
1	111	High-density urban fabric
2	112	Medium-density urban fabric
45	113	Low-density urban fabric
3	121	Industrial or commercial units
4	122	Road and rail networks and associated land
5	123	Port areas
6	124	Airports
7	131	Mineral extraction sites
8	132	Dump sites
9	133	Construction sites
10	141	Green urban areas
11	142	Sport and leisure facilities
12	211	Non-irrigated arable land
13	212	Permanently irrigated land
14	213	Rice fields
15	221	Vineyards
16	222	Fruit trees and berry plantations
17	223	Olive groves
18	231	Pastures
19	241	Annual crops associated with permanent crops
20	242	Complex cultivation patterns
21	243	Land principally occupied by agriculture, with significant areas of natural vegetation
22	244	Agro-forestry areas
23	311	Broad-leaved forest
24	312	Coniferous forest
25	313	Mixed forest
26	321	Natural grasslands
27	322	Moors and heathland
28	323	Sclerophyllous vegetation
29	324	Transitional woodland-shrub
30	331	Beaches, dunes, sands
31	332	Bare rocks
32	333	Sparsely vegetated areas
33	334	Burnt areas
34	335	Glaciers and perpetual snow
35	411	Inland marshes
36	412	Peat bogs
37	421	Salt marshes
38	422	Salines
39	423	Intertidal flats
40	511	Water courses
41	512	Water bodies
42	521	Coastal lagoons
43	522	Estuaries
44	523	Sea and ocean

Table 3. Nomenclature of CLC Refined 2006

For population disaggregation, we adopted the general methodological framework proposed by [8] and [9], while introducing some variations in different approaches. However, the following aspects were kept common to all the approaches:

- Reference population densities were obtained through empirical sampling of census tracts in LULC map, using the containment method (i.e., sampled census polygon must be totally within a single LULC patch or polygon)
- In the model, these relative densities are preserved within each population source zone (e.g., LAU2), considering its specific LULC composition

The basic modeling approach assumes that the GHSL layer mapping built-up percentage can be used to inform both the location and the density of built-up so as to support the disaggregation and spatial refinement of the distribution of population from administrative or enumeration areas to effective residential areas. In this framework, the use of GHSL as proxy for the distribution of residential population can potentially be augmented by including LULC maps as true ancillary data in the disaggregation process.

A generic model overview of the dasymetric mapping approach proposed, with options, is shown in Figure 4.

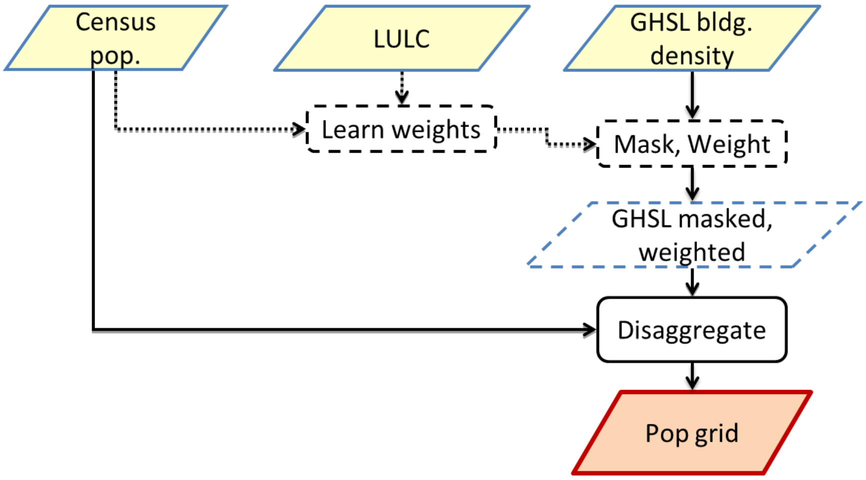


Figure 4. Workflow of generic model approach

Based on this framework, and considering the available input data sets, three main approaches are possible, each with potential variations in parameters:

A. No use of ancillary data (LULC) -- i.e. use of GHSL BU alone
B. Masking of GHSL with ancillary data (LULC)
C. Weighting of GHSL with ancillary data (LULC)

All approaches have advantages and shortcomings, which also depend on the characteristics and level of Census input data. For each approach, a brief presentation of the main challenges, as well as strengths (*Pros*) and weaknesses (*Cons*) follows.

A. No use of ancillary (LULC) data

Main challenges / decisions: none.

This approach is quite simple to implement, since it relies entirely on GHSL as proxy for resident population distribution. It allows changes after 2006 to be fully represented. The main drawback is the allocation of significant population to non-residential areas, due to absence of functional discrimination of buildings in GHSL.

Input Pop.	Pros	Cons
Tracts	Maximizes likelihood of volume preservation; Puts freedom and trust in GHSL; Allows changes after 2006	Some population possible in non-residential areas
LAU2/Km2	*(Idem as above)*	Significant population possible in non-residential areas

B. Masking of GHSL with ancillary (LULC) data

Main challenges / decisions:

- Which LULC classes to mask?
- How to select LULC classes to mask?

Input Pop.	Pros	Cons
Tracts		No preservation of volume assured; Population misplaced due to severe mismatch in scale, geometry and thematic content between CLC and census, GHSL
LAU2/Km2	Limits population allocated to non-residential areas (< Commission)	Limits changes after 2006; Destroys residential detection value of GHSL (> Omission)

While this approach may restrict allocation of population to non-residential areas, it limits representing changes after 2006 and confines the residential detection value of GHSL in rural areas if natural and agricultural classes are masked.

C. Weighting of GHSL with ancillary (LULC) data

Main challenges / decisions:

- Weight all classes?

- How to derive weights?
- Which weights to use when sampling not possible? (i.e., how to export to other areas)
- How significant are weights? (i.e., problem of under-sampled LULC classes)

Input Pop	Pros	Cons
Tracts	Maximizes likelihood of volume preservation; Preserves residential detection value of GHSL; Limits population going in non-residential Areas (< Commission); Allows changes after 2006	Population misplaced due to mismatch in scale, geometry and thematic between CLC and census, GHSL; Overall accuracy can be lower than approach B.
LAU2/Km2	*(Idem as above)*	No 'pure' sampling possible to derive weights

Application of IDM and this approach to population disaggregation with GHSL involves three main steps:

1. Combine GHSL and LULC layers (i.e., characterize GHSL cells in terms of their LULC class)
2. Using weights derived by empirical sampling, find the population that goes into each LULC class in each census source zone (i.e., target zone);
3. Disaggregate that population into GHSL cells, proportional to each cell's GHSL value

Quality Assessment and validation of population disaggregation methods

Many methods and metrics have been used to assess the accuracy of disaggregated population results, namely:

- Correlation coefficient (r)
- Root Mean Square Error (RMSE)
- Coefficient of Variance (CoV)
- Mean Absolute Error (MAE)
- Mean Absolute Percent Error (MAPE)
- Total Absolute Error (TAE)
- Relative Total Absolute Error (RTAE)

In the present work, for the purpose of comparing and benchmarking the results of using different disaggregation approaches, we adopted the same method as applied in [13] and [11], and using as metrics the Total Absolute Error (TAE) and the Relative Total Absolute Error (RTAE) as presented in those references. However it should be noted that these metrics are dependent on the data involved and especially on the spatial relation (scale and configuration) between source zones used for modeling and reference zones used for benchmarking. Therefore these measures are hardly comparable across different study areas and especially across modeling scales. In benchmarking results with 1 km grid, model values are rounded to integers.

Model development and experimentation was carried out on a (1) pilot study area in Portugal, (2) on continental Portugal, and (3) on a set of European countries: Italy (IT), United Kingdom (UK), Belgium (BE), Netherlands (NL), Slovenia (SI), Portugal (PT), and Spain (ES). Most of the tests were conducted with GHSL W7 (workflow 7), for comparability purposes.

Pilot study area in Portugal

To initially test the three approaches, a pilot study was implemented in Portugal. In Portugal all census levels are available: Blocks, Tracts, LAU2, LAU1, NUTS. The study area encompassed seven LAU2 (Local Area Units) administrative areas (23,504 ha) close to Lisbon, displaying a wide range of population densities and high LULC heterogeneity (presence of 21 of the 45 CLC06_r classes). Its resident population in 2011 was 129,752.

2011 census population was disaggregated from both census tracts and LAU2 (parishes) units using variations of the three main approaches outlined above, namely:

- Masking 1: this masking is more strict, masking GHSL in all CLC_r classes which are not expected to have a significant share of resident population; only six classes are not masked.
- Masking 2: this masking is less strict, allowing more freedom and trust in GHSL as proxy for resident population distribution. Only those CLC_r classes known to contain buildings or structures which are not residential are masked (6 classes).
- Weighting 1: this weighting scheme follows closely the recommendations of Mennis (2003) and Mennis and Hultgren (2006), using containment sampling to derive strict average population densities per CLC_r class, which are then summed and converted to ratios (0…1). While not all classes were sampled, a decision was made not to exclude (mask) any CLC_r class by attributing to those classes a weight lower than the lowest weight obtained for sampled classes (i.e., residual weight). The rationale for this option is two-fold: empirical

sampling revealed presence of resident population in almost all the CLC_r classes[2], even in unlikely ones; it maximizes likelihood of volume preservation by not excluding GHSL cells in CLC_r patches which may have been misclassified or suffer from generalization, especially when disaggregating from small source zones (i.e. census tracts).

- Weighting 2: this weighting scheme is a variation of the previous, as mitigation of the biases and distortions detected. It uses containment sampling to derive the share (%) of overall sampled population per CLC_r class, which are converted to ratios (0…1). While not all classes were sampled, a decision was made not to exclude (mask) any CLC_r class by attributing to those classes a weight lower than the lowest weight.

Modeling was carried out at 10 m and results were assessed using census blocks (the finest census zoning) as reference. For comparison, results from the simple gridded representation from the census using areal weighting were also included (Table 4).

					Complete Omission		Complete Commission	
Approach	**Method**	**Census input**	**TAE**	**RTAE**	**Blocks**	**Pop.**	**Blocks**	**Pop.**
Areal weighting	Only census	Tracts	84,871	0.65	0	0	100	14693
		LAU2	176,922	1.36	0	0	100	23895
A	Only GHSL	Tracts	75,662	0.58	0	0	100	12577
		LAU2	151,856	1.17	0	0	100	21341
B	Mask 1	Tracts	74,368	0.57	13	891	100	12194
		LAU2	130,896	1.01	13	891	100	12194
	Mask 2	Tracts	75,196	0.58	2	485	100	12303
		LAU2	138,137	1.06	2	485	100	22499
C	Weights 1	LAU2	122,819	0.95	0	0	100	20223
	Weights 2	LAU2	85,931	0.66	0	0	100	8244

Table 4. Benchmarking results of different approaches tested in pilot area in Portugal

Complete omission and commission errors are also important performance indicators of the disaggregation by showing for how many populated census units the model estimates zero population (omission), and the number of unpopulated census units for which the model estimates presence of population. Results show that a more restrictive masking (Mask 1) may lower the overall RTAE value and decrease Complete Commission error, but will increase Complete Omission by preventing population to be allocated to LULC zones where it exists.

Benchmarking results indicate that approach type C performs best by enabling to explore all the

[2] See Table 8 on p. 24

building-detection power of GHSL, and as such confirms GHSL as a good proxy for the location/distribution of residential population (Fig. 5). In the pilot study this approach has proven superior to disaggregating to ESSL masked by LULC, as implemented by Steinnocher et al. (2010) [10]. In particular, this approach has proven better at limiting major population re-allocation errors, such as population assigned to unpopulated areas (complete commission) and the converse (complete omission).

Benchmarking using census blocks (Fig. 6) shows that under-estimation of population occurs in small and dense blocks, while there is over-estimation in larger and lower density units.

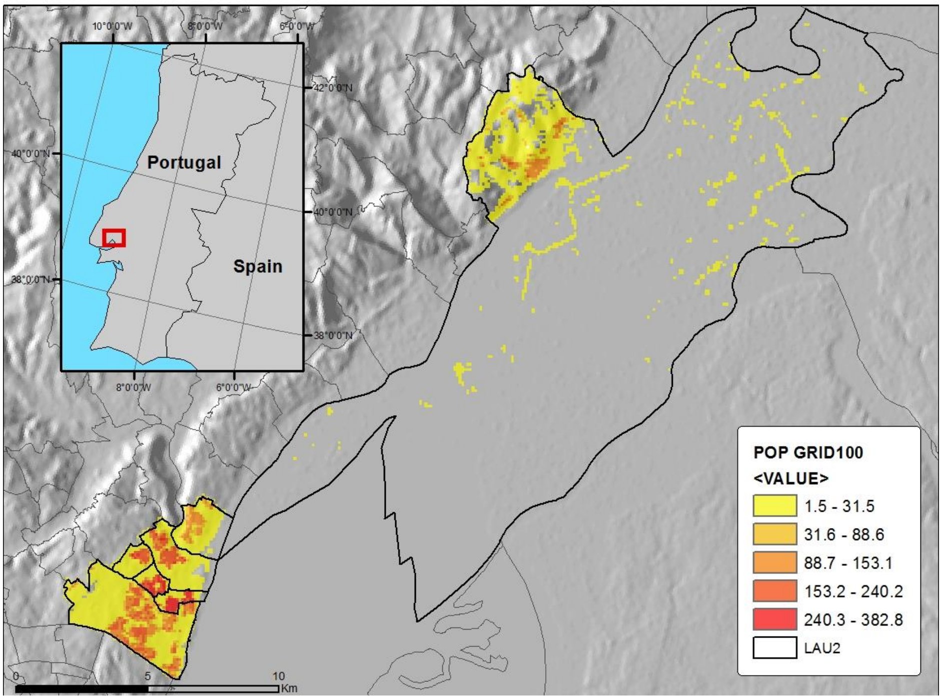


Figure 5. 100-m population count grid produced with approach type C. in pilot study area

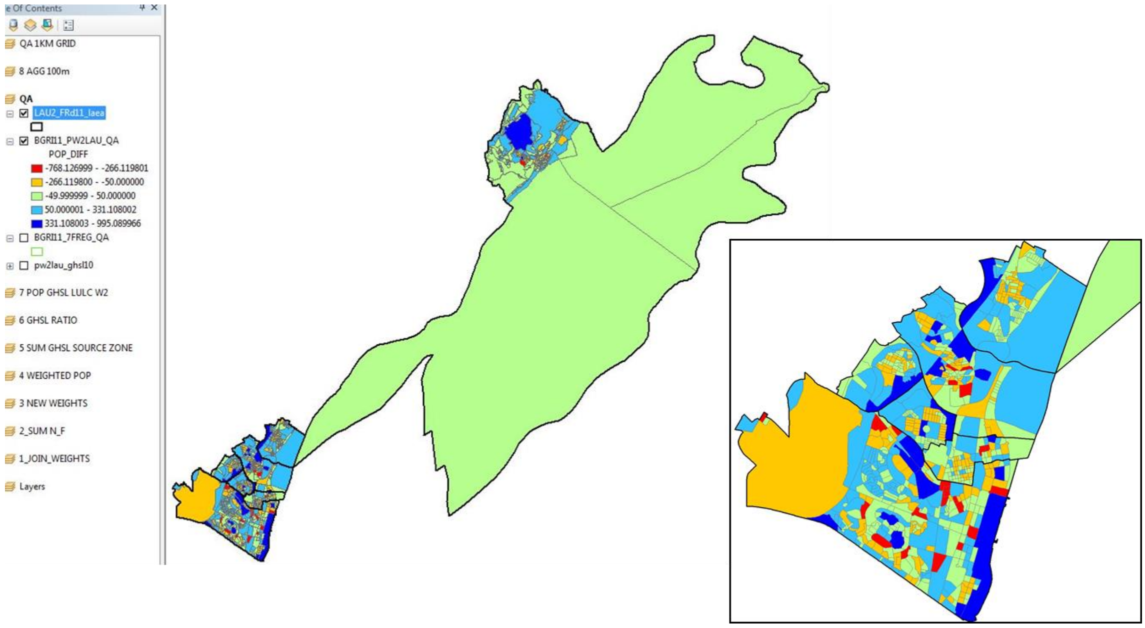


Figure 6. Benchmarking of approach C. using census blocks

Approach A has performed the worst, mostly due to error of commission, i.e., assignment of population to non-residential built-up areas.

Mainland Portugal

The series of systematic tests were then extended to the territory of continental Portugal (Population: 10,047,621). Census input data was used at the level of LAU2 (4,050 units), and data from GEOSTAT1B 1 km grid (bottom-up aggregation from NSO) was used as reference for benchmarking.

Five different approaches were tested and benchmarked, at 100 m resolution:

A. No use of ancillary data (LULC) - use of GHSL alone
B. Masking of GHSL with ancillary data (LULC)
C. Weighting of GHSL with ancillary data (LULC)
D. Weighting of GHSL with ancillary data (LULC), using total area of BU in LULC class to calculate class density
E. Approach D combined with masking some LULC classes

In approach D, the total area of built-up (from GHSL) in LULC class is used as denominator in the equation to calculate the class density. This approach requires availability of complete and final GHSL BU layer in order to calculate densities. Results (RTAE) are presented in Table 5 compared to the baseline representation possible by using census data alone: areal weighting of census population at LAU2 level.

Pop. surface	RTAE
Census LAU2	**0.82**
A	**0.46**
B	**0.44**
C	**0.35**
D	**0.32**
E	**0.31**

Table 5. Benchmarking results of five different approaches tested in Portugal

Test of approach type A (Disaggregating population from Census to GHSL BU) in Portugal shows major increase in the accuracy of population distribution, from 0.82 to 0.46, demonstrating that use of GHSL BU alone can improve over Census' population distribution. However, correlation analysis at LAU2 level between variables involved (Table 6) shows that correlation between census' resident population and number of buildings is not very strong. Correlation is stronger between buildings and GHSL, especially when the sum of built-up area is considered, indicating the added-value of this latter variable. Therefore approaches D and E were designed and tested, aiming at exploiting this information, while controlling for known issues.

r	Census (LAU2)			GHSL BU (10 m)	
Census (LAU2)	No. of residents	No. of buildings	No. of dwellings	No. of **GHSL** cells	Sum of **GHSL** values
No. of residents		**0.80**	0.97	**0.6**	0.65
No. of buildings			0.83	**0.82**	**0.85**
No. of dwellings				0.62	0.67

Table 6. Correlation matrix between census variables and GHSL BU

European countries

Census tract level (or infra-LAU2) shapefiles, circa 2011, were also provided by DG REGIO for a number of European countries: Scotland (SC), England (EN), Wales (WA), Northern Ireland (NI)[3], Netherlands (NL), Spain (ES), and Slovenia (SI). These data joined that for Portugal (PT) and Italy (IT), which were obtained through their respective national census portals (available tract-level data for Italy being provisional).

These data were used to develop a multi-country database of LU-population weights, by sampling census tracts within CLC06_r_v2 polygons (using containment sampling method). The Census tract data used for each country and respective sampling rates are presented in Table 7.

[3] These four countries collectively form the United Kingdom (UK)

					Sampled					
	Total polygons				Polygons		Area		Population	
Zone	No.	Tot. Ha	Mean Ha	Pop.	No.	%	Ha	%	Pop.	%
PT (continental)	255,844	8,908,213	34.8	10,047,621	69,137	27.0	113,804	1.3	2,372,452	23.6
IT	402,121	30,139,999	75.0	59,433,744	107,680	26.8	617,305	2.0	9,694,678	16.3
SC	46,351	7,793,711	168.1	5,295,403	21,480	46.3	47,484	0.6	2,310,269	43.6
NL	11,778	3,504,933	297.6	16,655,595	225	1.9	7,319	0.2	202,585	1.2
NI	4,537	1,414,907	311.9	1,810,863	467	10.3	3,237	0.2	143,178	7.9
EN + WA	181,408	15,128,077	83.4	56,075,912	50,244	27.7	258,250	1.7	15,128,403	27.0
ES	35,960	50,444,262	1,402.8	46,575,200	3,314	9.2	14,741	0.03	4,042,430	8.7
SI	*6,030*	*2,027,655*	*336.3*	*2,050,189*	*3*	*0.0*	*195*	*0.01*	*36*	*0.0*
TOTAL	937,999	117,334,102	125.1	195,894,338	252,547	26.9	1,062,140	0.9	33,893,995	17.3

Table 7. Countries and respective censuses used to develop a database of LU weights

Table 7 shows that overall, 17% of the total population was sampled, corresponding to 27% of the census tract polygons but only to 1% of the area. This confirms that this sampling method favors the selection of small, higher population density polygons. The sampling rate varied significantly among countries, with this value being very low in the Netherlands (due to the combination of relatively large census polygons and a fragmented landscape as captured by CLC06_r_v2 – see Fig. 7).

Figure 7. Sampled census tract units (blue) completely within CLC06_r_v2 polygons, in Amsterdam

In Slovenia the sampling rate was so low (three polygons) that samples were discarded.
Table 8 summarizes the overall results of empirical sampling, showing for each LULC class the number of census polygons sampled and respective area, population, population density (Pop/ha) and density fraction (*Df*).

		Polygons						
CLC_r Code	CLC Code	No.	%	Ha	Pop.	%	Pop/ha	*Df(*100)*
0		236	0.09	18	29	0.0001	1.6	0.23
1	111	47,291	18.73	42,732	8,384,276	24.7368	196.2	28.75
2	112	132,421	52.43	368,118	23,097,258	68.1456	62.7	9.19
3	121	4,315	1.71	13,162	356,170	1.0508	27.1	3.96
4	122	201	0.08	157	6,492	0.0192	41.4	6.06
5	123	286	0.11	377	21,262	0.0627	56.4	8.26
6	124	15	0.01	648	1,509	0.0045	2.3	0.34
7	131	64	0.03	199	7,167	0.0211	36.0	5.27
8	132	4	0.00	1	53	0.0002	49.1	7.20
9	133	295	0.12	687	37,668	0.1111	54.8	8.03
10	141	248	0.10	277	14,689	0.0433	53.1	7.78
11	142	929	0.37	2,711	13,490	0.0398	5.0	0.73
12	211	13,800	5.46	382,039	424,947	1.2538	1.1	0.16
13	212	250	0.10	3,638	3,593	0.0106	1.0	0.14
14	213	475	0.19	30,106	5,419	0.0160	0.2	0.03
15	221	1,970	0.78	11,769	42,172	0.1244	3.6	0.53
16	222	1,004	0.40	4,989	24,270	0.0716	4.9	0.71
17	223	2,154	0.85	11,598	66,878	0.1973	5.8	0.84
18	231	1,361	0.54	10,126	46,749	0.1379	4.6	0.68
19	241	5,963	2.36	20,140	95,742	0.2825	4.8	0.70
20	242	11,557	4.58	41,248	276,487	0.8157	6.7	0.98
21	243	6,874	2.72	16,023	142,403	0.4201	8.9	1.30
22	244	148	0.06	1,964	1,619	0.0048	0.8	0.12
23	311	2,625	1.04	46,111	49,758	0.1468	1.1	0.16
24	312	529	0.21	5,868	4,137	0.0122	0.7	0.10
25	313	716	0.28	3,015	10,024	0.0296	3.3	0.49
26	321	247	0.10	6,805	2,534	0.0075	0.4	0.05
27	322	40	0.02	268	486	0.0014	1.8	0.27
28	323	157	0.06	4,415	2,718	0.0080	0.6	0.09
29	324	1,062	0.42	4,183	12,132	0.0358	2.9	0.43
30	331	19	0.01	40	16	0.0000	0.4	0.06
31	332	93	0.04	340	50	0.0001	0.1	0.02
32	333	75	0.03	251	135	0.0004	0.5	0.08
33	334	6	0.00	22	21	0.0001	1.0	0.14
34	335							
35	411	7	0.00	186	94	0.0003	0.5	0.07
36	412							
37	421	7	0.00	38	92	0.0003	2.4	0.35
38	422	1	0.00	7	35	0.0001	5.4	0.78
39	423							
40	511	34	0.01	66	36	0.0001	0.5	0.08
41	512	39	0.02	874	674	0.0020	0.8	0.11
42	521	56	0.02	3,786	199	0.0006	0.1	0.01
43	522							
44	523							
45	113	14,973	5.93	23,138	740,512	2.1848	32.0	4.69
TOT.		252,547	100	1,062,140	33,893,995	100.0	31.9	100
							682.5	

Table 8. Overall results of empirical sampling using census tract data for PT, ES, IT, UK, and NL

The empirical sampling revealed presence of enumerated residential population in 40 of the 45 classes of the CLC06_r_v2 layer, plus in the 0-NoData zones (Table 9). This is mostly due to the generalization implied by the large MMU of CLC06_r_v2 and the geometric inconsistencies between this layer and large-scale census data. These issues also contribute to the existence of spurious densities for some LULC classes, especially those suffering from low sampling rates or that are heavily artificialized (e.g., classes 4, 5, 7-10).

The sampling has also shown that some classes contain a significant share (%) of the total population sampled, albeit with low densities (e.g. class 12), and vice-versa. For these reasons an analysis was made with the double objective of:

(1) identifying LULC classes with similar behavior regarding the distribution of resident population (to be aggregated or merged) and

(2) detecting *problematic* LULC classes (to be masked).

In order to combine the information on the share (%) of the total sampled population and its density for each class, a new metric was calculated by multiplying these two values (column C in Table X below). LULC classes were then ranked according to this metric, and following this ranking, nine class groups were defined. However, since the ranking of artificial non-residential classes is affected by the problem of spurious densities, these classes were arbitrarily assigned to a new class, taking into consideration the CLC/CLC_r class definitions and knowledge gained during the tests conducted.

			A	B		C	D
CLC_r Code	**CLC Code**	**Pop.**	**%**	**Pop/ha**	***Df(*100)***	***[A*B]***	**New class**
1	111	8,384,276	24.7368	196.2	28.75	4853.4688	1
2	112	23,097,258	68.1456	62.7	9.19	4275.7343	2
45	113	740,512	2.1848	32.0	4.69	69.9225	3
3	121	356,170	1.0508	27.1	3.96	28.4356	4
9	133	37,668	0.1111	54.8	8.03	6.0907	4
20	242	276,487	0.8157	6.7	0.98	5.4680	6
21	243	142,403	0.4201	8.9	1.30	3.7341	6
5	123	21,262	0.0627	56.4	8.26	3.5367	5
10	141	14,689	0.0433	53.1	7.78	2.3018	5
12	211	424,947	1.2538	1.1	0.16	1.3946	6
19	241	95,742	0.2825	4.8	0.70	1.3429	6
17	223	66,878	0.1973	5.8	0.84	1.1378	7
4	122	6,492	0.0192	41.4	6.06	0.7928	5
7	131	7,167	0.0211	36.0	5.27	0.7608	5
18	231	46,749	0.1379	4.6	0.68	0.6368	7

			A	B		C	D
CLC_r Code	CLC Code	Pop.	%	Pop/ha	*Df(*100)*	*[A*B]*	New class
15	221	42,172	0.1244	3.6	0.53	0.4459	7
16	222	24,270	0.0716	4.9	0.71	0.3484	7
11	142	13,490	0.0398	5.0	0.73	0.1981	5
23	311	49,758	0.1468	1.1	0.16	0.1584	8
29	324	12,132	0.0358	2.9	0.43	0.1038	8
25	313	10,024	0.0296	3.3	0.49	0.0983	8
13	212	3,593	0.0106	1.0	0.14	0.0105	8
6	124	1,509	0.0045	2.3	0.34	0.0104	5
24	312	4,137	0.0122	0.7	0.10	0.0086	8
8	132	53	0.0002	49.1	7.20	0.0077	5
28	323	2,718	0.0080	0.6	0.09	0.0049	8
22	244	1,619	0.0048	0.8	0.12	0.0039	8
14	213	5,419	0.0160	0.2	0.03	0.0029	8
26	321	2,534	0.0075	0.4	0.05	0.0028	8
27	322	486	0.0014	1.8	0.27	0.0026	9
41	512	674	0.0020	0.8	0.11	0.0015	9
37	421	92	0.0003	2.4	0.35	0.0006	9
38	422	35	0.0001	5.4	0.78	0.0006	9
32	333	135	0.0004	0.5	0.08	0.0002	9
35	411	94	0.0003	0.5	0.07	0.0001	9
0	0	29	0.0001	1.6	0.23	0.0001	9
33	334	21	0.0001	1.0	0.14	0.0001	9
40	511	36	0.0001	0.5	0.08	0.0001	9
42	521	199	0.0006	0.1	0.01	0.0000	9
31	332	50	0.0001	0.1	0.02	0.0000	9
30	331	16	0.0000	0.4	0.06	0.0000	9
34	335					0.0000	9
36	412					0.0000	9
39	423					0.0000	9
43	522					0.0000	9
44	523					0.0000	9
TOT.		33,893,995	100.0	31.9			
				682.5			

Table 9. Ranking of CLC_r classes according to the new metric (column C) and new classes proposed

Identification of LULC classes for masking (i.e., density fraction set to zero) was based on the empirical sampling conducted (sampling rates), the ranking proposed, and analysis of CLC/CLC_r class definitions and knowledge gained during the tests conducted. Results of this exercise are presented in Table 10.

		A	B	C	D
CLCr Code	CLC Code	Bldgs.	Resid.	Mask	New class
0				Y	9
1	111	Y	Y		1
2	112	Y	Y		2
3	121	Y	Y		4
4	122	Y		Y	5
5	123	Y		Y	5
6	124	Y		Y	5
7	131			Y	5
8	132			Y	5
9	133	Y	Y		4
10	141	Y		Y	5
11	142	Y	Y		5
12	211				6
13	212				8
14	213				8
15	221				7
16	222				7
17	223				7
18	231				7
19	241				6
20	242	Y	Y		6
21	243				6
22	244				8
23	311				8
24	312				8
25	313				8
26	321				8
27	322			Y	9
28	323				8
29	324				8
30	331			Y	9
31	332			Y	9
32	333			Y	9
33	334			Y	9
34	335			Y	9
35	411			Y	9
36	412			Y	9
37	421			Y	9
38	422			Y	9
39	423			Y	9
40	511			Y	9
41	512			Y	9
42	521			Y	9
43	522			Y	9
44	523			Y	9
45	113	Y	Y		3

Table 10. CLC_r classes suggested for masking and merging

Table 10 shows that the CLC/CLC_r class definitions allow for significant presence of buildings in 11 of the 45 classes (column A), and in seven of those classes a substantial part of those buildings will be residential (column B). However, in four of the 11 classes a substantial presence of residential buildings *is not* to be expected (e.g., port areas, airports). Since the GHSL layer used detects all buildings, regardless of their use, it is especially important to mask the latter group of classes, as to limit significant re-allocation of resident population into non-residential areas. The remaining classes suggested for masking are those for which, based on the empirical sampling, there is very weak evidence of presence of resident population (i.e., less than 1,000 people).

The empirical sampling using census tract data for PT, ES, IT, UK, and NL has also revealed substantial differences in the landscape (as represented by CLC06_r_v2) and consequently in the LULC-Population relationships within Southern-Mediterranean countries (PT, ES, IT) (Table 11) and in the Central-Northern European ones (UK, NL) (Table 12).

		Polygons						
CLC_r Code	CLC Code	No.	%	Ha	Pop.	%	Pop/ha	*Df(*100)*
0		236	0.13	18	29	0.000	1.59	0.26
1	111	45,727	25.39	39,756	8,031,013	49.852	202.01	32.48
2	112	63,655	35.34	88,193	6,095,674	37.839	69.12	11.11
3	121	3,888	2.16	11,748	256,172	1.590	21.81	3.51
4	122	198	0.11	144	5,728	0.036	39.78	6.40
5	123	223	0.12	168	9,276	0.058	55.38	8.90
6	124	13	0.01	642	1,363	0.008	2.12	0.34
7	131	30	0.02	85	646	0.004	7.64	1.23
8	132	4	0.00	1	53	0.000	49.14	7.90
9	133	176	0.10	261	11,753	0.073	45.02	7.24
10	141	156	0.09	114	2,537	0.016	22.30	3.58
11	142	908	0.50	2,630	10,490	0.065	3.99	0.64
12	211	13,643	7.57	367,812	397,830	2.470	1.08	0.17
13	212	250	0.14	3,638	3,593	0.022	0.99	0.16
14	213	475	0.26	30,106	5,419	0.034	0.18	0.03
15	221	1,970	1.09	11,769	42,172	0.262	3.58	0.58
16	222	1,004	0.56	4,989	24,270	0.151	4.87	0.78
17	223	2,154	1.20	11,598	66,878	0.415	5.77	0.93
18	231	1,236	0.69	2,568	30,061	0.187	11.71	1.88
19	241	5,963	3.31	20,140	95,742	0.594	4.75	0.76

		Polygons						
CLC_r Code	CLC Code	No.	%	Ha	Pop.	%	Pop/ha	*Df(*100)*
20	242	11,553	6.41	41,128	276,217	1.715	6.72	1.08
21	243	6,871	3.81	16,009	142,066	0.882	8.87	1.43
22	244	148	0.08	1,964	1,619	0.010	0.82	0.13
23	311	2,620	1.45	46,096	49,354	0.306	1.07	0.17
24	312	529	0.29	5,868	4,137	0.026	0.70	0.11
25	313	715	0.40	2,881	9,761	0.061	3.39	0.54
26	321	242	0.13	3,798	1,459	0.009	0.38	0.06
27	322	38	0.02	259	269	0.002	1.04	0.17
28	323	157	0.09	4,415	2,718	0.017	0.62	0.10
29	324	1,062	0.59	4,183	12,132	0.075	2.90	0.47
30	331	19	0.01	40	16	0.000	0.40	0.06
31	332	93	0.05	340	50	0.000	0.15	0.02
32	333	75	0.04	251	135	0.001	0.54	0.09
33	334	6	0.00	22	21	0.000	0.97	0.16
34	335							
35	411	6	0.00	14	34	0.000	2.48	0.40
36	412							
37	421	7	0.00	38	92	0.001	2.39	0.38
38	422	1	0.00	7	35	0.000	5.36	0.86
39	423							
40	511	34	0.02	66	36	0.000	0.54	0.09
41	512	36	0.02	781	199	0.001	0.25	0.04
42	521	56	0.03	3,786	199	0.001	0.05	0.01
43	522							
44	523							
45	113	13,954	7.75	17,526	518,312	3.217	29.57	4.75
TOT.		180,131	100.00	745,850	16,109,560	100	21.60	100

Table 11. Aggregated results of empirical sampling for *Southern* (PT, ES, IT) group of countries

		Polygons						
CLC_r Code	CLC Code	No.	%	Ha	Pop.	%	Pop/ha	*Df(*100)*
0								
1	111	1,564	2.16	2,976	353,263	1.99	118.69	15.87
2	112	68,766	94.96	279,926	17,001,584	95.60	60.74	8.12
3	121	427	0.59	1,415	99,998	0.56	70.69	9.45
4	122	3	0.00	13	764	0.00	59.40	7.94
5	123	63	0.09	210	11,986	0.07	57.18	7.65

		Polygons						
CLC_r Code	CLC Code	No.	%	Ha	Pop.	%	Pop/ha	*Df(*100)*
6	124	2	0.00	6	146	0.00	23.50	3.14
7	131	34	0.05	115	6,521	0.04	56.86	7.60
8	132							
9	133	119	0.16	426	25,915	0.15	60.80	8.13
10	141	92	0.13	163	12,152	0.07	74.64	9.98
11	142	21	0.03	81	3,000	0.02	37.10	4.96
12	211	157	0.22	14,227	27,117	0.15	1.91	0.25
13	212							
14	213							
15	221							
16	222							
17	223							
18	231	125	0.17	7,558	16,688	0.09	2.21	0.30
19	241							
20	242	4	0.01	120	270	0.00	2.26	0.30
21	243	3	0.00	14	337	0.00	24.29	3.25
22	244							
23	311	5	0.01	15	404	0.00	26.61	3.56
24	312							
25	313	1	0.00	134	263	0.00	1.97	0.26
26	321	5	0.01	3,007	1,075	0.01	0.36	0.05
27	322	2	0.00	9	217	0.00	23.65	3.16
28	323							
29	324							
30	331							
31	332							
32	333							
33	334							
34	335							
35	411	1	0.00	172	60	0.00	0.35	0.05
36	412							
37	421							
38	422							
39	423							
40	511							
41	512	3	0.00	93	475	0.00	5.10	0.68
42	521							

		Polygons						
CLC_r Code	**CLC Code**	**No.**	**%**	**Ha**	**Pop.**	**%**	**Pop/ha**	***Df(*100)***
43	522							
44	523							
45	113	1,019	1.41	5,611	222,200	1.25	39.60	5.29
TOT.		72,416	100	316,290	17,784,435	100	56.23	100

Table 12. Aggregated results of empirical sampling for *Northern* (UK, NL) group of countries

The sampling of UK and NL indicates that there is significant presence of residential population on only six CLC_r classes, and 99% of the sampled population is found in the three Urban fabric classes (i.e., 1, 2, 45).

Testing in five countries: BE, NL, SI, PT, ES

Based on the above information, it was decided to develop different sets of weights and to test the disaggregation from LAU2 at 100 m in five countries (BE, NL, SI, PT, ES) for which existed reference data for benchmarking. This reference data was obtained from the GEOSTAT1B project's 1 km grid, which contained bottom-up aggregated counts for these countries.

Table 13 shows the characteristics of population input units and reference layer used for BE, NL, SI, PT, and ES.

		COUNTRIES				
		BE	**NL**	**SI**	**PT**	**ES**
LAU2	No.	589	418	210	4,050	8,026
	TOT. POP. 2011:	10,951,266	16,574,989	2,050,189	10,047,621	44,568,515
	Mean POP	18,593	39,653	9,762	2,481	5,553
	Mean Ha	5,207	8,497	9,656	2,200	6,193
REFERENCE (1 Km grid)	TOT. POP.:	10,989,886	16,640,942	2,048,647	9,923,171	44,562,241
	Min	0	0	0	0	0
	Max	29,100	23,485	10,504	21,823	53,119
	No. of 1 km cells	31,532	36,991	20,996	87,987	501,819

Table 13. Characterization of input units and reference layer for BE, NL, SI, PT, ES

By considering three groups of countries and two approaches to derive the reference population densities by LULC class, six different sets of weights were developed and tested (Table 14). The three groups of countries considered are:

- The five countries (PT, ES, IT, UK, and NL), noted as *Global 5C*

- The three Southern countries (PT, ES, IT), noted as *South 3C*
- The two countries (UK and NL), noted as *North 2C*

The two methods used to calculate the reference population densities by LULC class were:

- Dividing the total sampled population from census by the total area of BU in the LULC class, noted as method *3B*
- Dividing the total sampled population from census by the total area of the sampled census polygons (strict density) and then heuristically adjusting the densities and masking some classes, noted as method *1H* – these weights are presented on page 35.

Results of experimenting these weight sets for population disaggregation on the five test countries (BE, NL, SI, PT, ES) are shown in Table 14.

		COUNTRIES				
	WEIGHT SET	**BE**	**NL**	**SI**	**PT**	**ES**
	TAE – Tot. Absolute Error (pax)					
1	C-specific				3,021,447	
2	Global 5C 3B	4,957,661	5,510,634			
3	Global 5C 1H	3,591,333	5,625,505	1,027,568	3,304,934	18,117,659
4	South 3C 3B	4,862,948	6,542,333			
5	South 3C 1H	3,692,311	5,511,102	981,837	3,259,180	17,912,634
6	North 2C 3B	6,014,676	6,624,178			
7	North 2C 1H	3,163,263	4,926,952	1,151,202	3,684,751	18,704,512
	RTAE - TAE relative to country population					
1	C-specific				0.304	
2	Global 5C 3B	0.45	0.331			
3	Global 5C 1H	0.33	0.34	0.50	0.333	0.41
4	South 3C 3B	0.44	0.39			
5	South 3C 1H	0.34	0.33	0.48	0.328	0.40
6	North 2C 3B	0.55	0.40			
7	North 2C 1H	0.29	0.30	0.56	0.37	0.42

Table 14. Results of benchmarking for BE, NL, SI, PT, ES

Benchmarking results show that heuristic weight sets produce disaggregations having the lowest error, with the South set (3C1H) producing the best grids in three countries (SI, PT, ES), including the southern ones, and the North set (2C1H) performing best for BE and NL. Results for Portugal show that performance of heuristic South set (3C1H) is not too far from the best result obtained with country-specific weights (C-specific).

However, the Global set (5C1H) appears to constitute a good compromise if only a weight set is to be used, since it is the worst-performing only in NL.

Using Urban Atlas 2012 as land use data

For certain areas and cities, Urban Atlas 2012 (UA2012) was made available as a more updated, spatially-detailed source of land use geoinformation. UA2012 maps land use/cover as polygons using a nomenclature with 25 classes and an official minimum mapping unit (MMU) of 0.25 ha for Artificial surfaces (and 1 ha for remaining classes) [14].

Due to the very small MMU of UA, no direct empirical sampling (with containment method) is possible to derive specific reference population densities for its classes. Still, due to its thematic proximity to CLC, weights derived for the latter can be transmitted to UA, as long as a reclassification table is provided. However, correspondence between CLC06r and UA is not direct and therefore cannot be obtained automatically by simply truncating UA class codes.

A thematic integration of CLC06r and UA classification systems was investigated and a reclassification table proposed (Table 15).

Urban Atlas		CLC Refined 2006		
CODE2012	**ITEM2012**	**CLC_R**	**CLC_CODE**	**CLC06_R**
11100	Continuous Urban Fabric (S.L. > 80%)	1	111	High-density urban fabric
11210	Discontinuous Dense Urban Fabric (S.L. : 50% - 80%)	2	112	Medium-density urban fabric
11220	Discontinuous Medium Density Urban Fabric (S.L. : 30% - 50%)	2	112	Medium-density urban fabric
11230	Discontinuous Low Density Urban Fabric (S.L. : 10% - 30%)	45	113	Low-density urban fabric
11240	Discontinuous Very Low Density Urban Fabric (S.L. < 10%)	45	113	Low-density urban fabric
11300	Isolated Structures	45	113	Low-density urban fabric
12100	Industrial, commercial, public, military and private units	3	121	Industrial or commercial units
12210	Fast transit roads and associated land	4	122	Road and rail networks and associated land
12220	Other roads and associated land	4	122	Road and rail networks and associated land
12230	Railways and associated land	4	122	Road and rail networks and associated land

Urban Atlas		CLC Refined 2006		
CODE2012	**ITEM2012**	**CLC_R**	**CLC_CODE**	**CLC06_R**
12300	Port areas	5	123	Port areas
12400	Airports	6	124	Airports
13100	Mineral extraction and dump sites	7	131	Mineral extraction sites
13300	Construction sites	9	133	Construction sites
13400	Land without current use	8	132	Dump sites
14100	Green urban areas	10	141	Green urban areas
14200	Sports and leisure facilities	11	142	Sport and leisure facilities

Table 15. Proposed reclassification table to integrate thematically CLC06r and UA maps

UA class '13400' (Land without current use) has no direct correspondence in CLC at level 3 and represents well the challenges of LU reclassification. Depending on the criterion considered, this class can be made to correspond *correctly* to three different CLC classes:

- 141-Green urban areas: if priority is placed on the **land cover**, since these plots are usually covered with overgrown vegetation (includes urban agriculture in some cities);
- 133-Construction sites: if primacy is placed on the **land use dynamics**, since these plots are typically building land and will tend to be built up sooner or later;
- 132-Dump sites: if priority is placed on the **land use**, since these plots are normally considered having a marginal use; this was selected as best option in the context of our purpose.

The UA is converted from vector to raster using the Maximum Combined Area (MCA) algorithm[4].

The result of reclassification of UA2012 into CLC06r classes in the City of Luxemburg is illustrated in Fig. 8.

[4] If there is more than one feature in a cell with the same value, the areas of these features will be combined. The combined feature with the largest area within the cell will determine the value to assign to the cell.

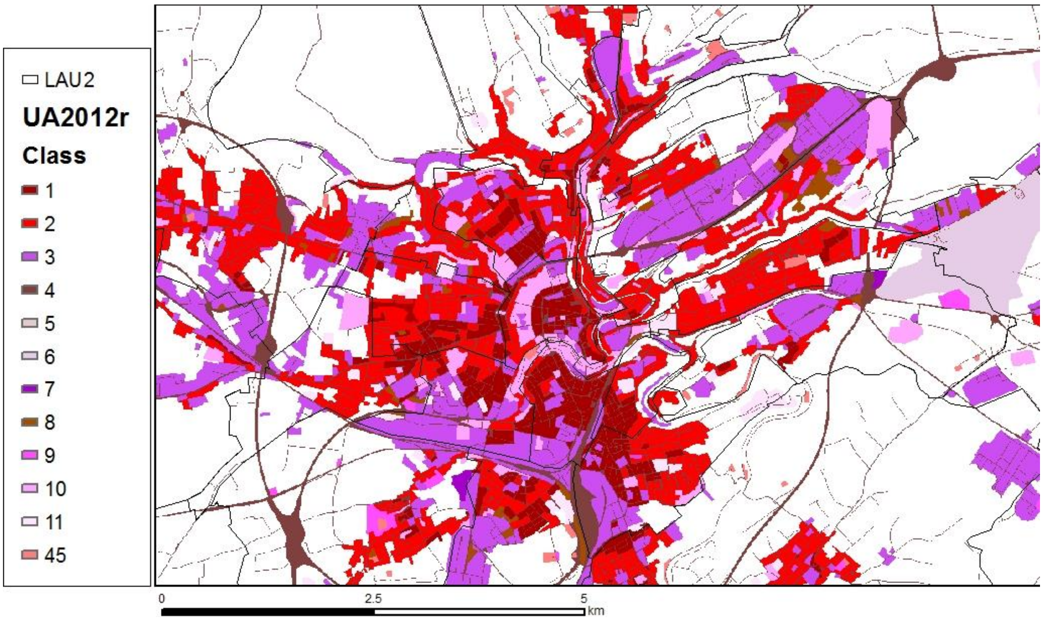


Figure 8. Result of reclassifying UA2012 into CLC06r classes in LU

Adopted methodology and tools

As stated above, the base methodology proposed follows the IDM approach by Mennis (2003), and Mennis and Hultgren (2006). However, to obtain reference population densities by LULC class (*Df*), a semi-empirical approach was adopted:

- *Df* were defined using the empirical base recommended by Mennis and Hultgren, 2006 (i.e., considering the total area of sampled census tracts in the LULC class)
- These densities were heuristically tuned and re-defined, especially for problematic classes
- In CLC06_r classes for which there was not sufficient evidence of strong LU-Population association, weight was set to zero (i.e., masked)

The model is fully automated in ArcGIS, based on Python programming language, and includes a new tool in ArcToolbox (Fig. 9).

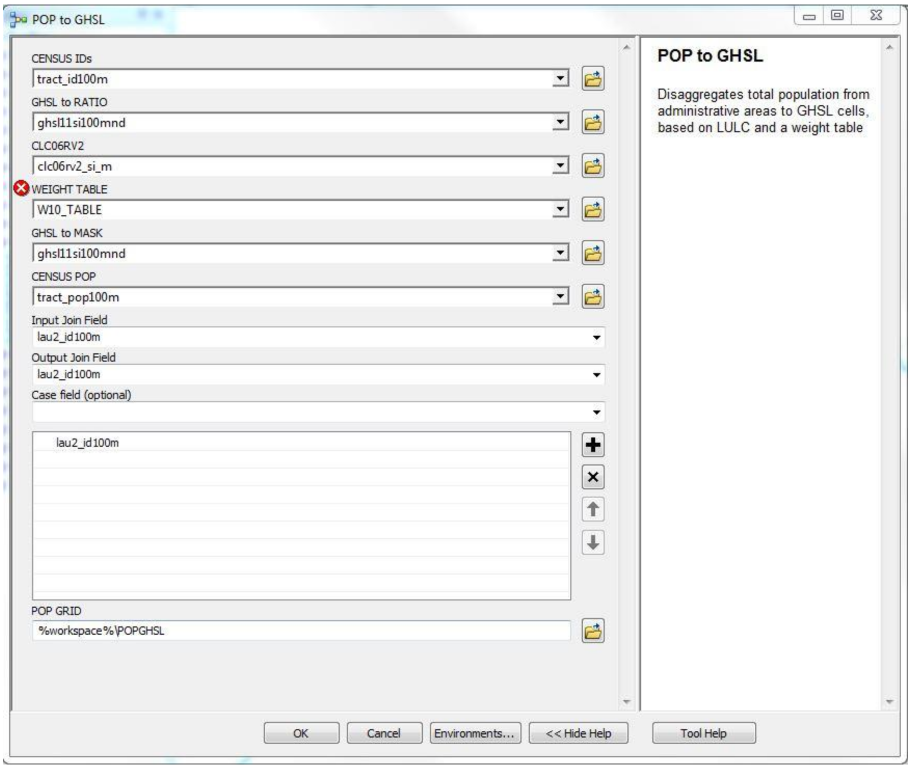


Figure 9. GUI of developed tool in ArcToolbox for population disaggregation

The model operates on raster data and requires the following variables as inputs: population source zones (population counts and zone IDs), the GHSL BU grid, a LULC map, and a table storing weights for each LULC class.

The modeling steps are detailed in Fig. 10:

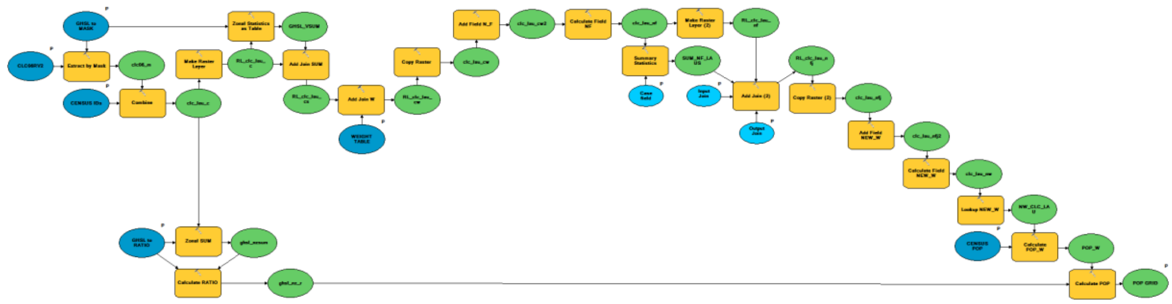


Figure 10. Sequence of steps in Model Builder for population disaggregation

Quality assessment (QA) of output grids was also automated through a model and respective new tool in ArcToolbox (Fig. 11). This tool allows calculating TAE and RTAE for the produced population grid, requiring as input a polygon layer containing reference (i.e., ground truth) population values, to be compared with modeled values.

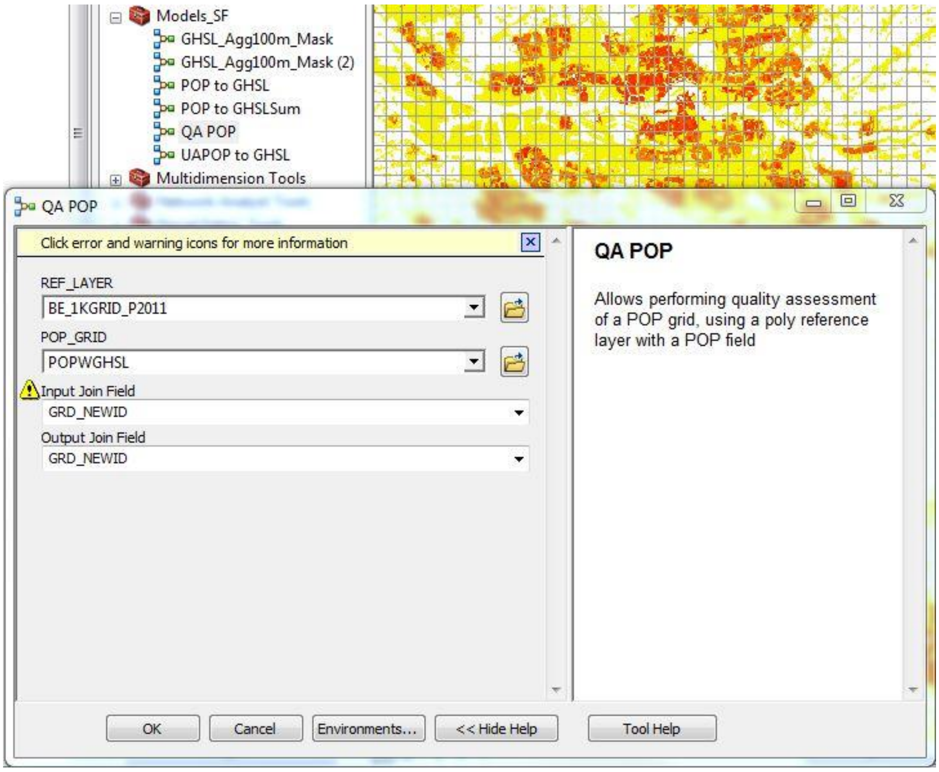


Figure 11. GUI of developed tool in ArcToolbox for quality assessment of population grids

The proposed weights (density and density fractions) for population disaggregation with CLC06_r_v2 are presented in Table 16:

		GLOBAL		SOUTH		NORTH	
CLC_r Code	CLC Code	Pop/ha	WG1H	Pop/ha	WS1H	Pop/ha	WN1H
0	0						
1	111	180.0	57.58	200.0	58.65	120.0	52.17
2	112	60.0	19.19	70.0	20.53	60.0	26.09
45	113	30.0	9.60	30.0	8.80	40.0	17.39
3	121	10.0	3.20	10.0	2.93	5.0	2.17
4	122						
5	123						
6	124						
7	131						
8	132						
9	133	10.0	3.20	4.0	1.17		
10	141						
11	142	4.0	1.28	3.0	0.88		
12	211	2.0	0.64	1.0	0.29	3.0	1.30
13	212	0.1	0.03	0.1	0.03		
14	213	0.1	0.03	0.1	0.03		
15	221	1.0	0.32	1.0	0.29		
16	222	1.0	0.32	1.0	0.29		
17	223	1.0	0.32	1.0	0.29		
18	231	1.0	0.32	3.0	0.88	2.0	0.87
19	241	3.0	0.96	4.0	1.17		
20	242	3.0	0.96	5.0	1.47		
21	243	3.0	0.96	5.0	1.47		
22	244	0.1	0.03	0.1	0.03		
23	311	0.1	0.03	0.1	0.03		
24	312	0.1	0.03	0.1	0.03		
25	313	0.1	0.03	0.3	0.09		
26	321	0.1	0.03	0.1	0.03		
27	322						
28	323	0.1	0.03	0.1	0.03		
29	324	0.1	0.03	0.3	0.09		
30	331						
31	332						
32	333						
33	334						
34	335						
35	411						
36	412						
37	421						
38	422						
39	423						
40	511						
41	512						
42	521						
43	522						
44	523						

Table 16. Proposed weights for population disaggregation with CLC06_r_v2

In brief, the proposed methodology ensures the following:

- Volume preservation (preserve census population totals)
- Approaches a functional definition of population (closest to ‘residence-based’ population)
- Is apt for using different census levels as input
- Takes advantage of and preserves, as much as possible, the GHSL detection of built-up

Production and results

1 km population grids were so far produced for Luxemburg (LU), Cyprus (CY), and Malta (MT), with production also planned for Greece (EL) and possibly Italy (IT). Due to the data available and characteristics of these countries, for production purposes different LULC layers and sets of weights were used (Table 17).

	COUNTRY			
Input dataset	**LU**	**CY**	**MT**	***EL***
Population	LAU2+	LAU2	LAU2	*LAU2*
GHSL BU	GHSL10w11	GHSL100w11	GHSL10w11	*GHSL100w11*
LU	UA2012	CCL06rv2	UA2006(09)	*CLC2000*
Weights	WN1H	WS1H	WS1H	*WS2H*

Table 17. Input datasets available and used for each country

For countries where the UA is used as source of LULC information, only four classes are considered as target for population disaggregation: 1, 2, 3, and 45. The rationale is that the scale and MMU of UA allows for capturing virtually all residential built-up areas in these classes, without the spatial generalization effects observed in CLC06r.

For Malta, it was decided to use UA2006 as source of land use information (instead of CLC06rv2) due to the fact that CLC06rv2 showed severe misregistration with other spatial layers (Figure 12).

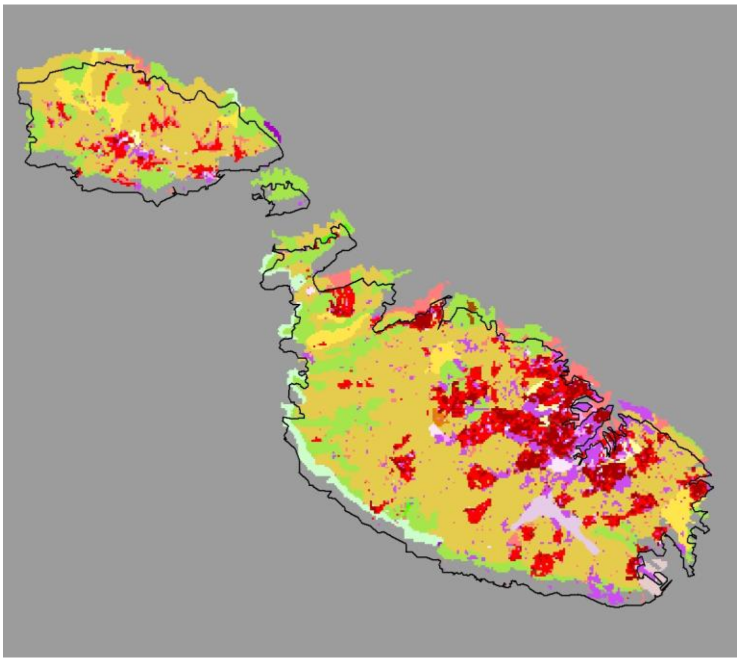


Figure 12. Misregistration of CLC06_r_v2 in Malta

Figure 13 shows the1 km population grid for LU in 2011.

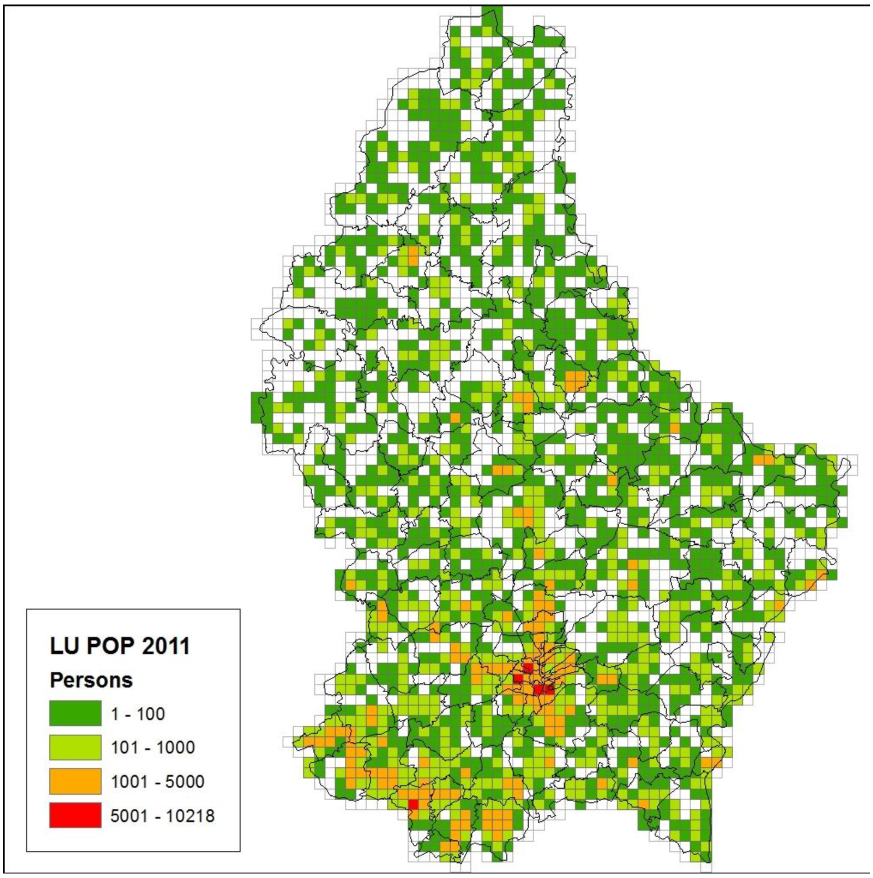


Figure 13. 1 km population grid for LU in 2011

Figure 14 shows the 1 km population grid for MT in 2011.

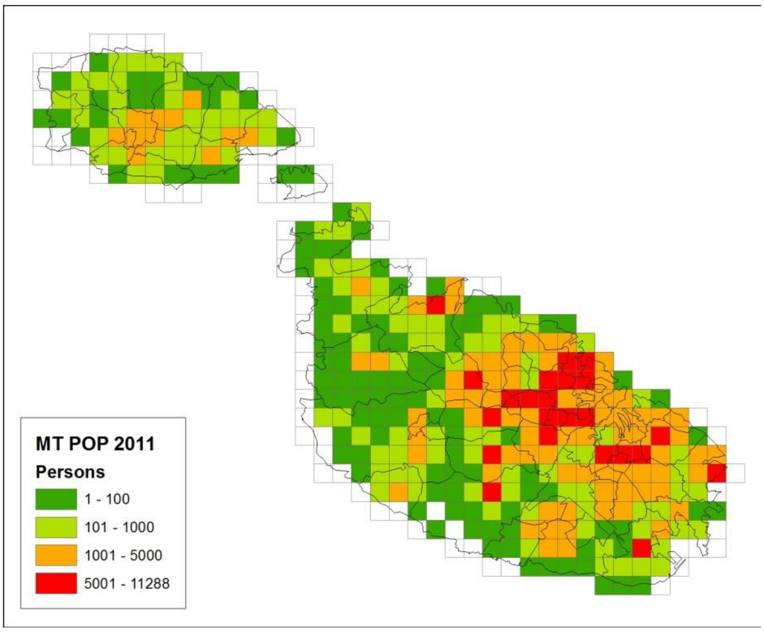


Figure 14. 1 km population grid for MT in 2011

Figure 15 shows the 1 km population grid for CY in 2011.

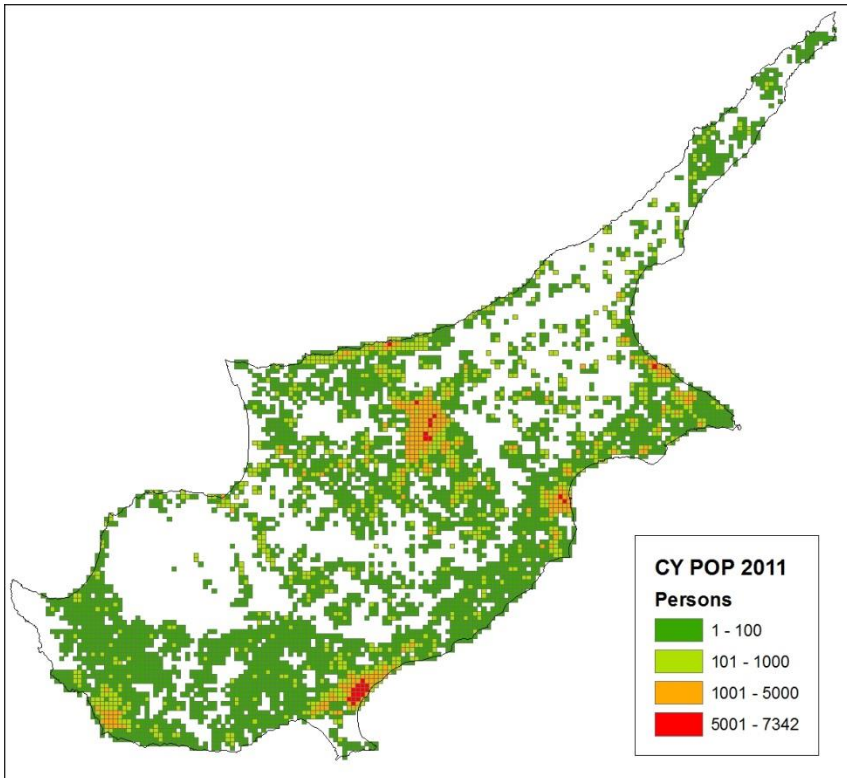


Figure 15. 1 km population grid for CY in 2011

Quality control of results was carried out as much as possible. This has included visual inspection of output and assurance that partial and total counts were preserved in the disaggregation process, by re-aggregating results to source zones using zonal statistics.

Conclusions and outlook

Based on model development and testing it is possible to conclude:

- All models using GHSL improved on census representation of population distribution and results were superior to other existing approaches
- GHSL BU constitutes a good proxy, under certain conditions, for residential population distribution at high resolution, informing both on its location and specific densities
- Ancillary data on LULC, even if coarser, improves results, especially when LULC classes are weighted

Some remaining challenges and potential solutions include:

- Under-estimation of occupancy of apartment buildings – requires information on building height (number of floors)
- Coarse LULC, mix of LU and LC – requires wide-coverage, detailed land use at compatible scales, and detachment of LU from LC classifications

Acknowledgements

We thank Filipe Batista e Silva for fruitful discussions during the course of this work.

Bibliography

[1] Pesaresi M. and Ehrlich D. (2009) A methodology to quantify built-up structures from optical VHR imagery, in Global Mapping of Human Settlement Experiences, Datasets, and Prospects, P. Gamba and M. Herold, Eds. Boca Raton, FL, USA: CRC Press, ch. 3, pp. 27–58.

[2] Ferri, S., Syrris, V., Florczyk, A., Scavazzon, M., Halkia, S. and Pesaresi M. (2014) A new map of the European settlements by automatic classification of 2.5-m resolution SPOT data, Proceedings of IGARSS 2014, Quebec, Canada, 13-18 July 2014.

[3] Freire, S. (2010). Modeling of spatio-temporal distribution of urban population at high-resolution – value for risk assessment and emergency management. In: Konecny, M., Zlatanova, S., Bandrova, T.L. (eds.), Geographic Information and Cartography for Risk and Crisis Management. Lecture Notes in Geoinformation and Cartography. Springer Berlin Heidelberg, pp. 53-67.

[4] Wright J. K. (1936). A method of mapping densities of population, The Geographical Review, 26(1), 103–110.

[5] Eicher C.L., Brewer C.A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartogr Geogr Inf Sci 28:125–138.

[6] Wu, S-S., Qiu, X., & Wang, L. (2005). Population estimation methods in GIS and remote sensing: A review. GIScience and Remote Sensing, 42, 80–96.

[7] Langford, M. (2013). An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis. 45, 324-344.

[8] Mennis, J. (2003). Generating surface models of population using dasymetric mapping. The professional Geographer, 55(1), 31–42.

[9] Mennis, J. and Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. Cartogr. Geogr. Inf. Sci., 33,179–194.

[10] Steinnocher K., Kaminger I., Köstl M. and Weichselbaum J., (2010). Gridded Population – new data sets for an improved disaggregation approach. e- proceedings of the 3th European Forum for Geostatistics Conference, 5-7 October 2010, Tallinn, Estonia.

[11] Batista e Silva, F., Gallego J. & Lavalle C. (2013). A high-resolution population grid map for Europe. Journal of Maps, 9:1, 16-28.

[12] Batista e Silva, F., Lavalle, C., & Koomen, E. (2012). A procedure to obtain a refined European land use/cover map. Journal of Land Use Science, 8(3): 255-283.

[13] Gallego, F. J., Batista, F., Rocha, C., & Mubareka, S. (2011). Disaggregating population density of the European Union with CORINE Land Cover. International Journal of Geographical

Information Science, 25(12), 2051–2069.

[14] European Union, 2011. Mapping Guide for a European Urban Atlas. Available at: http://www.eea.europa.eu/data-and-maps/data/urban-atlas/#tab-methodology

European Commission
EUR 27482 – Joint Research Centre – Institute for the Protection and Security of the Citizen (IPSC)

Title: Urban and Regional Built-up Analysis (URBA)

Author(s): Sergio FREIRE, Stamatia HALKIA, Martino PESARESI, Daniele EHRLICH

Luxembourg: Publications Office of the European Union

2015 – 46 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online), ISSN 1018-5593 (print)

ISBN 978-92-79-52068-6 (PDF)
ISBN 978-92-79-52067-9 (print)

doi:10.2788/68127

LB-NA-27482-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:10.2788/68127

ISBN 978-92-79-52068-6

Publications Office