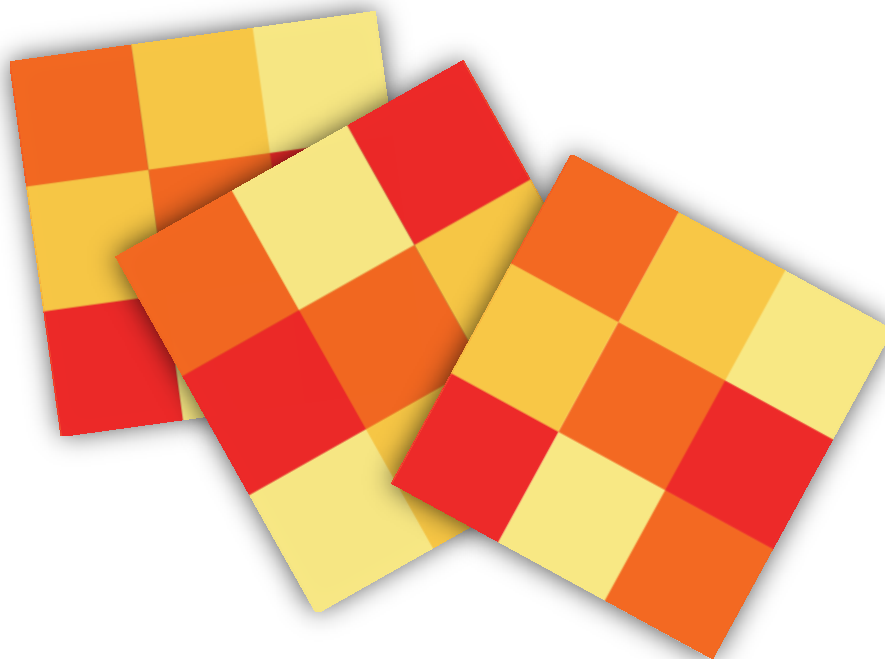


BULGARIAN POPULATION GRID 2011

(Geostatistics by mixed methods)



Project	ESSnet project GEOSTAT 1B	
Agreement No.	50502.2009.004-2011.536	
WP	1B: Geostatistics by disaggregation or mixed methods	
Task	1B.3: The production of datasets according to the “Hybrid” Approach	
Deliverable	Technical report, WP1B.3 Dataset	
Date	2014-02-10	
Contributors	<National Statistical Institute of the Republic of Bulgaria>	< Arslan Ahmedov; Irena Dudova>

ABSTRACT

The aim of this paper is to introduce the process behind creation of the Bulgarian population grid 2011 with using combination of “bottom-up” (aggregation) and “top-down” (disaggregation) approaches. The aggregation approach is applied mainly in bigger settlements where more than half of the population is concentrated and where the share of geocoded address data is very high. The disaggregation (dasymetric binary method) is applied where the data about exact location of the population is not geocoded, mainly rural areas with small settlements without officially adopted cadastral plans. The paper also provides detailed information about the design of the methodology and results of the prototype of Bulgarian grid map.

Sofia, Friday, February 14, 2014

AUTHORS

1. AHMEDOV, Arslan
2. DUDOVA, Irena

Experts in department of Geostatistics, Demography and Social Statistics Directorate

National Statistical Institute of the Republic of Bulgaria

E-mails: Aahmedov@nsi.bg; Idudova@nsi.bg

TABLE OF CONTENTS

Bulgarian Population Grid 2011	1
Geostatistics by mixed methods	1
Abstract	2
Authors	2
Table of contents	3
Introduction	4
Specify and Design	5
Collect and Harmonize	6
Build geocoding framework	7
Generate population grid	9
Analyze	11
Disseminate and use	13
Data specifications	13
Further	13
Maintenance and Sustainability	13
References	14
APPENDIX I	15
APPENDIX II	16
APPENDIX III	17

INTRODUCTION

The Bulgarian population grid 2011 was produced by National Statistical Institute of the Republic of Bulgaria (BNSI) as a co-partner in the ESSnet project: "Geostat 1B – representing Census data in European population grid." Bulgaria as a country without national population grid dataset and traditions in producing grid-based statistics was involved in Work Package 1B of the project: Geostatistics by disaggregation or mixed methods.

The main task was to produce population distribution for Bulgaria per 1km² grid net - INSPIRE compliant with data from Census 2011.

For implementation of the task a mix of point-based and polygon-based georeferences was used. For aggregation phase as a beginning of population grid generation process a dataset with address points mass was collected. This address dataset is gathered from different sources – cadaster address points where available, cadaster parcel's center points with address description, address locations pinpointed on map and geocoded addresses to the precision of the address.

The decision of using hybrid solution was driven by several reasons:

- Census 2011 is not geocoded;
- The lack of national register of addresses and buildings with coordinates;
- Aggregation cannot be applied to the whole territory of the country;
- To improve the accuracy of population density distribution by getting a higher quality than just disaggregation.

This technical report aims to describe the used methodology, analyze and share results and finally to marks some basic conclusion. The structure of the document will follow the main phases of the production process settled in the beginning of the project. (Figure 1)

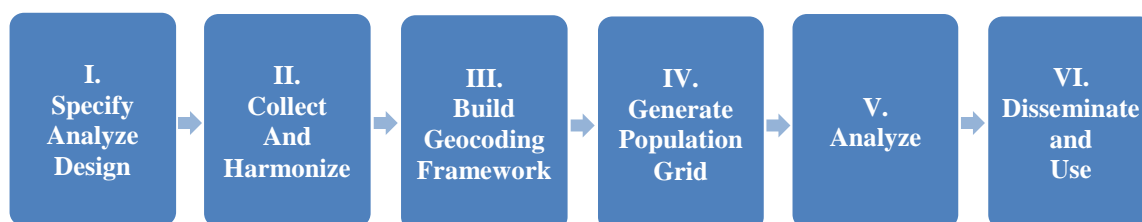


Figure 1: Production process flowchart for creating Bulgarian population grid

SPECIFY AND DESIGN

The first and the most important phase of the production: to specify, analyze the problem and to design the generic framework for creating the population grid. During the project a couple of examples of used hybrid solutions from other countries (Poland, Portugal, France, Spain and Ireland) were studied. The conclusion after comparing each case showed that there is no universal solution that can be applied. Even more, it is difficult to give a general overview of a “hybrid”, it can be sensed in every part – in the statistical datasets, in the methods, in the spatial datasets.

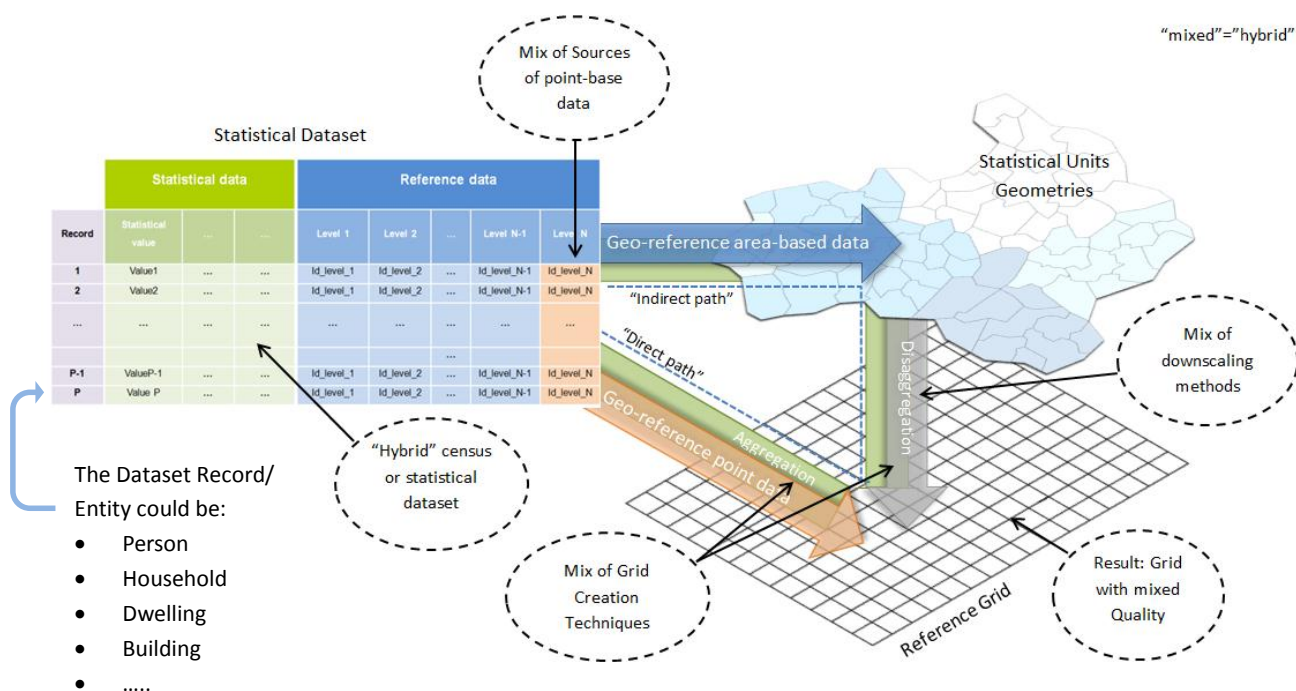


Figure 2: Sense of “Hybrid”

Ideally a geo-reference for each entity/ record in Statistical Dataset should be:

- Longitude, latitude (X,Y-position of the entity)
- A grid unit reference (grid cell identifier)

For Bulgaria the mentioned two cases are not an option taking in consideration the available data. That's why additional sources of coordinate data for the entities have been used. In the Census 2011 the geographical position of each entity is based on physical address (the most common case for many countries). Consequently to solve the problem of linking the statistics to location, Geocoded Address Framework have been established.

The way of the statistical data to the grid is via proxy geometries, point-based for a “direct path” and area-based for an “indirect path”. The link between statistical dataset and map features is released by common identifier codes. These codes are in a reference information part of the statistical dataset. First, statistical figures are assigned to the corresponding map features geometries (*source* data) and then certain disaggregation techniques (informed by *ancillary* data) take part in the process of transforming to grid geometry representation (*target*).

COLLECT AND HARMONIZE

The data, of the above mentioned categories, used to accomplish the project task:

Data category	Dataset	National coverage	Spatial reference	Time period	Provider
Source	Census 2011	Full	N/A	2011	BNSI
	Points of the addresses	Partial	1970; Sofia 1950	2010+/-1	Geodesy, Cartography and Cadastre Agency
	Centroids of parcels	Partial	1970	2010	Geodesy, Cartography and Cadastre Agency
	Geocoded and pinpointed addresses	Partial	WGS84	2011	BNSI
	Census EAs	Sofia; Yambol	1970; Sofia 1950	2011	BNSI
	Localization Units	Sofia; Varna; Plovdiv; Burgas	WGS84 UTM35N	2011	BNSI
	Administrative Territorial units	Full	WGS84 UTM35N	2011	Ministry of Agriculture and Foods
Target	Grid 1sq.km	Full	ETRS89 LAEA	-	Eurostat, (downloaded from www.efgs.info)
Ancillary	Land cover and land use dataset of populated and build-up areas	Partial (95%)	WGS84 UTM35N	2010 +/- 1	Ministry of Agriculture and Foods
	Urban Atlas	Partial	ETRS89 LAEA	2006	European Environment Agency (EEA)

Table 1: Used datasets

By using different sources of coordinate data the coverage of point-based data was improved. Spatial transformations were undertaken for harmonization of different datasets. The spatial datasets were transformed to the ETRS89 LAEA reference system. The phase of collection and harmonization of all needed datasets was the most *time* and *effort* consuming part of the project. The time of producing grid is essential in order to be updated regularly in the future.

BUILD GEOCODING FRAMEWORK

The mixed solution uses point-based and area-based geographical units as proxies for population distribution. The coded address of the building in the census dataset serves as a multilevel address locator and links population data with corresponding proxy geometry.

Determining Geography Population Entities that Statistical Data refers to




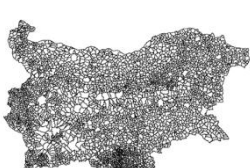
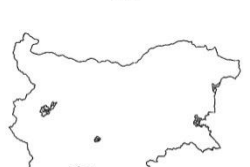

Statistical Units Dataset Hierarchy		Identifiers code list	Spatial representation of Statistical Units / Map features	Resolution
Level 1	Country	<BG>		MIDDLE
Level 2	Districts	UCATTU <Districts code>		
Level 3	Municipalities	UCATTU <Municipalities code>		
Level 4	Populated places/ settlements	UCPP <Populated places codes>		
Level 5	Localization Units ¹	CRA <LU codes>		HIGH
Level 6	Addresses	<Address codes>		

Figure 4: Census Geocoding Framework

Statistical Dataset is commonly produced with reference to different geographical levels.

Statistical data should refer to statistical unit through their identifier / code.

¹ **Localization units** in a given populated place: streets, boulevards, squares, dwelling complexes, neighborhoods, villa areas.

Figure 4 shows Statistical Units organized into hierarchical structure, associated with lower units. Upper (middle resolution) levels are usually sub-national (administrative) geographical divisions that compose tessellation and are traditionally used for reporting statistical data and are therefore well incorporated into statistics production process. High resolution Levels, in other hand, are still not very well interoperable with statistical data.

Incompleteness –

- Map features of lower levels had usually incomplete coverage
- References to high resolution levels are missing for some records in statistical dataset.

(Statistics produced in non-census years does not keep track of such references.) Incompleteness in lower levels leads to applying downscaling techniques to distribute the corresponding population.

Note*: Only localization units of a type “Dwelling complex” in the four biggest towns are delineated. The resident population in these settlements is 28.2% from total population of the country.

Localization units of a type “Dwelling complex” are very densely populated neighborhoods. The point is to localize people more precisely and to overcome the problems when downscaling such an “islands” with very high density of the population and at the same time with the low density of buildings and low values of imperviousness. This leads to a big underestimation of the population counts in this geographical unit and overestimation of nearby areas.

GENERATE POPULATION GRID

The transfer of the population counts from the source to the target geometry is done by means of GIS operations (in the ArcGIS environment). The infographic describes in details the order of the steps and operations done.

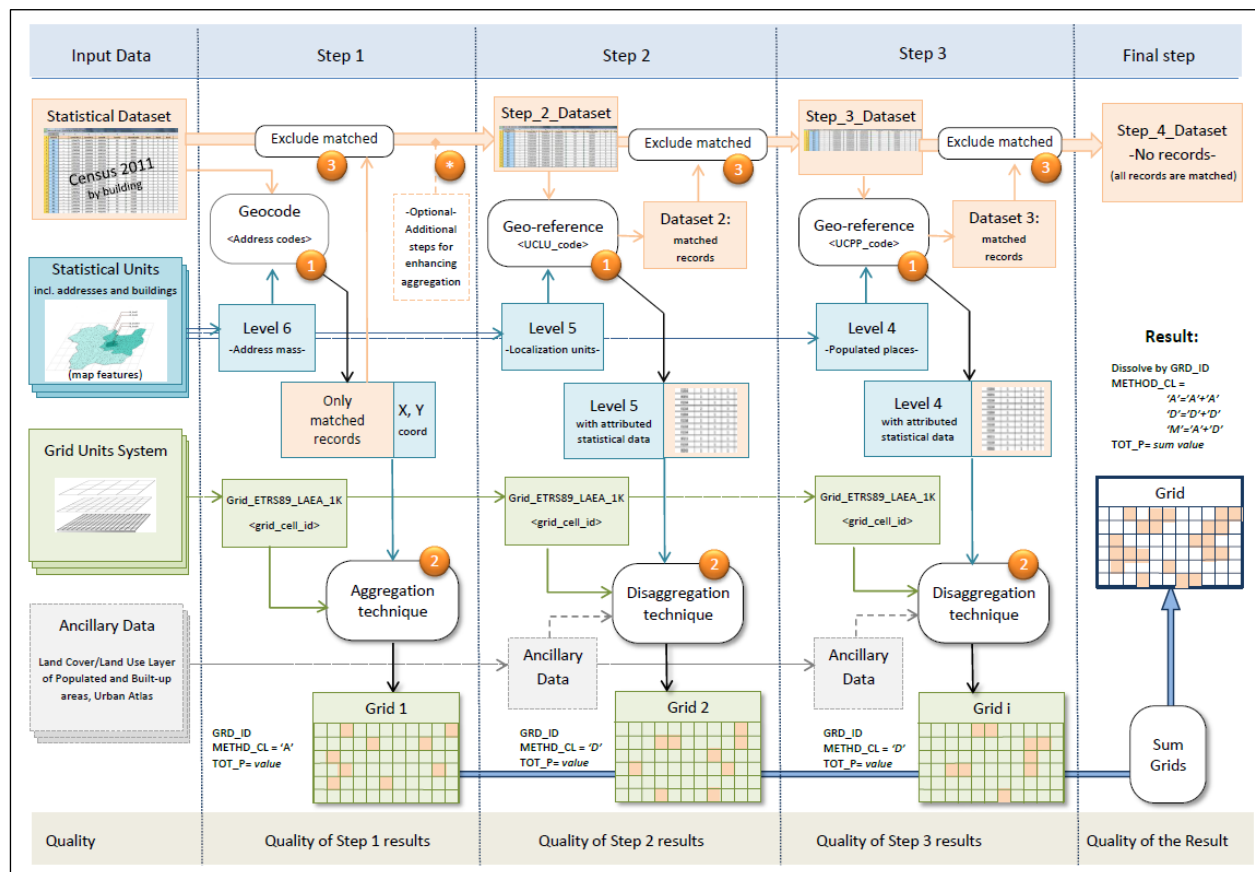


Figure 6: Process for generation population grid (see Appendix I)

The production process is documented per grid cell.

- 1 Point 1** in every step is where the Source geometries with population data attributed are formed by assigning the statistical population data to the map features.
- 2 Point 2** The gridding techniques
- 3 Point 3** The already distributed population is removed from the statistical dataset and the reduced one is passed to the next step, providing that no population can be distributed twice.
- * Point *** the Enumeration Areas (EA) for towns of Sofia and Yambol were used to enhance aggregations results. EA that fit inside the target grids without splitting. (For details see reference 3. Corcoran, D. (2011))

Aggregation:

Spatial join of point features and grid cells applied. (For details see reference 2. EFGS (2012))

Disaggregation:

Herein a common and well-known algorithm for areal interpolation is used - the Binary dasymetric technique. The method is based on ancillary data and divides source zones on populated and unpopulated parts. The population is distributed only to populated zones. The disaggregation in this example is performed in two scales of source zones – Level 5 (available Localization units), Level 4 (Populated places).

Assumptions:

Population and sub-population totals within a given geographic unit (source zone) are assumed to be distributed evenly. The population density is estimated for every source zone.

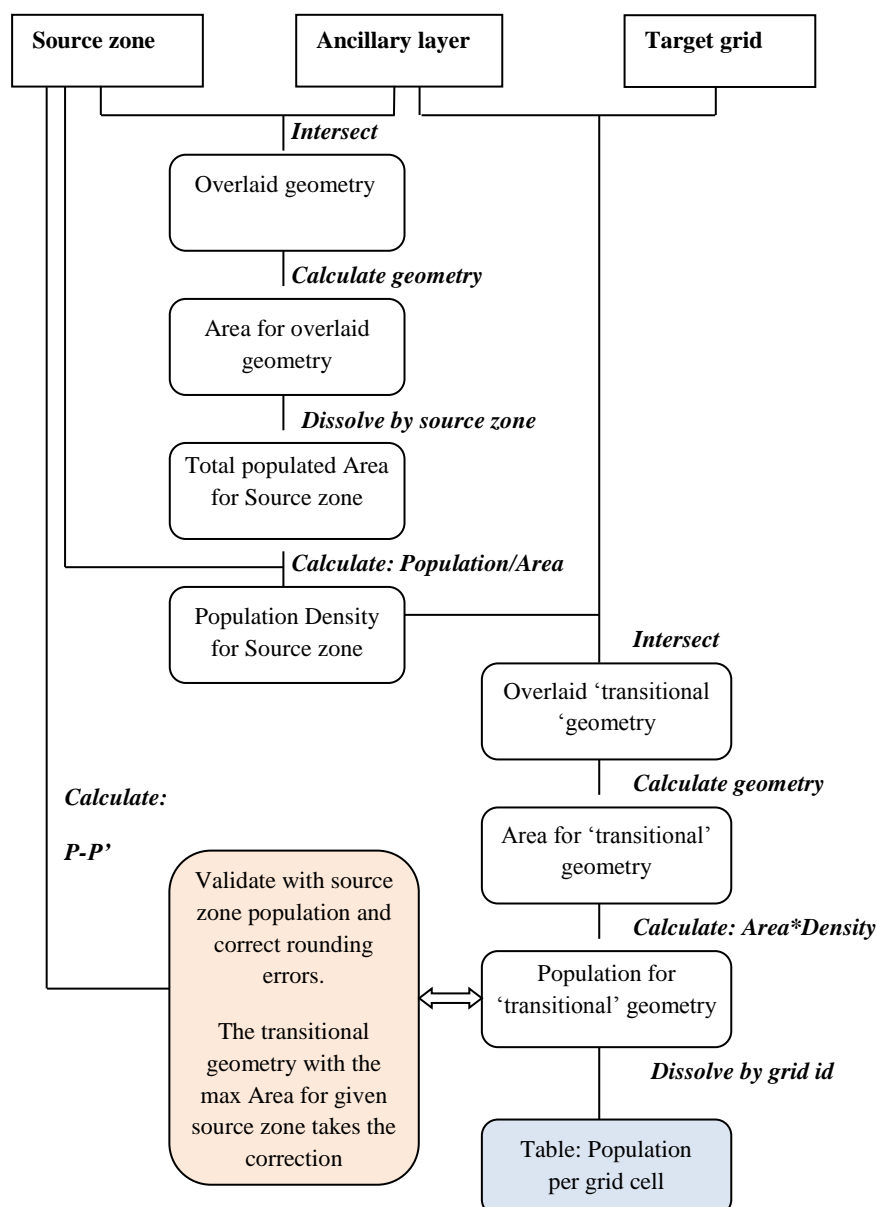


Figure 7: Disaggregation algorithm

The Final step is to sum-up grids from all previous steps.

(Dissolve by GRID_ID; METHD_CL = {A, D, M} where 'A' = 'A'+ 'A'; 'D' = 'D'+ 'D'; 'M' = 'A'+ 'D')

ANALYZE

Bulgaria has population resident and concentrated in smaller or bigger populated places and not scattered all over the territory thus a good behavior of the dasymetric binary method for distribution from detailed settlement areas layer was expected. Also the bigger settlements have geocoded address data, so the disadvantage of the binary method in these areas is not so feasible.

	Territorial Units Populated places		Population		% population distributed by aggregation	% area covered
	count	Structure - %	count	Structure - %	%	%
Total	5302	100.0	7364570	100.0	57.3	100
0	181	3.4	0	0.0	0	16.0
1 - 99	1627	30.7	58979	0.8	4.2	
100 - 499	1939	36.6	501828	6.8	5.2	51.2
500 - 999	753	14.2	531184	7.2	8.0	
1 000 - 4 999	665	12.5	1260320	17.1	20.1	26.8
5 000 - 9 999	60	1.1	425892	5.8	43.4	
10 000 - 19 999	33	0.6	446214	6.1	35.4	3.6
20 000 - 49 999	25	0.5	792874	10.8	67.0	
50 000 - 99 999	12	0.2	876356	11.9	92.5	2.3
100 000 +	7	0.1	2470923	33.6	89.5	

Table 2: Population breakdown by territorial units and rate of population distribution by aggregation

A total number of 113388 grids cover the surface area of Bulgaria (111 002 km²), 23618 (20.8%) of them are inhabited.

Population per grid	Number of grids	Grid population	% Population aggregated
1-4	3264	8513	5.7%
5-19	5128	54336	4.1%
20-199	10046	778650	8.1%
200-499	2820	888854	14.3%
500-4999	2054	2676026	48.5%
5000+	306	2958191	92.2%
Total	23618	7364570	57.3%

Table 4: Inhabited grids

So we use different approaches to grid statistical data. In fact, hybrid approach of producing population grid is in combining grids produced with different approaches and techniques. The main question is “hybrid” quality. How to measure the result-grid quality? One of the answers may be - the more (X, Y) data for point-based proxy geometries, the better quality of the result. That’s why it is good to use every possible data source for geocoding of the entities to coordinates, that is costly reasonable (quality, time, efforts).

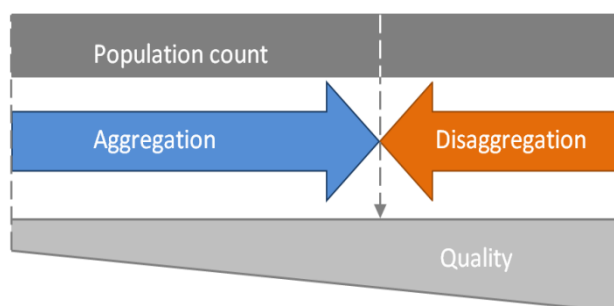


Figure 8: Quality

Method	Population	
Aggregation	4219321	57.3%
Disaggregation	3145249	42.7%
Total	7364570	100.0%

Table 3: Population breakdown by the type of used method

Map of the Population grid 2011 of Bulgaria (See Appendix II)

Map of the Population grid 2011 of Bulgaria by method (See Appendix III)

DISSEMINATE AND USE

The results of the project will be disseminated free of charge via the BNSI and the EFGS websites as a package of statistical data together with a grid net shape file, an INSPIRE metadata and documentation related to the files.

Data specifications

- Grid data for GEOSTAT dataset is referenced to the European grid Grid_ETRS89-LAEA_1K
1 sq. km grid size
- INSPIRE grid coding system (Data specifications on Geographical Grid System)
- Confidentiality rules are not applied for the absolute numbers of population count at this scale.
- No restrictions are applied to publishing total numbers, and other variables – population breakdowns such as age groups and sex.
- Data update frequency- not fixed at the moment.
- Data quality measures – will be documented in INSPIRE metadata in abstract and lineage fields

Further

1. Improvement of the quality:
 - Location of population in smaller areas. Development of the localization units level map features;
 - Improvement of disaggregation: More accurate downscaling techniques (Statistical modeling of the density by class of current land use layer is possible but needs further analysis). Other sources of detailed spatial data besides national urban land cover/land use layer (SSL, GHSL, OSM, etc.)
2. Promote the use of grid statistics.

Maintenance and Sustainability

Update procedure for the data sources: population data, spatial data

The population data are updated continuously and are sustained with high resolution between censuses. The impossibility to produce the grid entirely with bottom-up approach comes from the lack of exhaustive national point-based spatial layer describing buildings or addresses. Population and demographic data are annually published in a specialized publications.

Information System Demography ISD, maintained by the BNSI is the main source of statistical data on population, demographic events and migration. In the Population census years, the ISD is updated through the census data. The census information is used as a basis, updated afterwards with demographic events occurred during the years between the censuses.

Better quality in address part of the population micro data are expected. The amount of point-based information also is expected to grow.

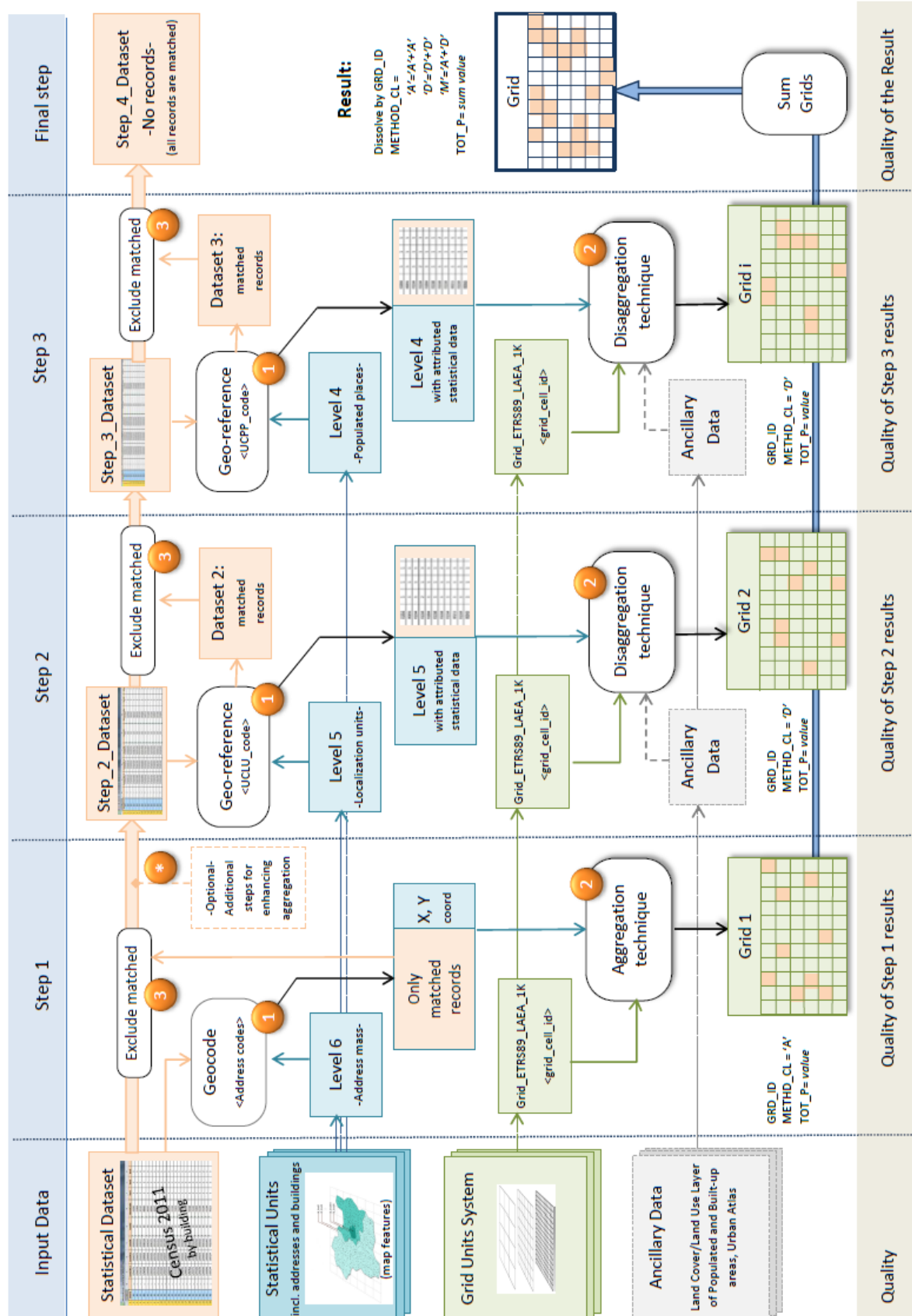
REFERENCES:

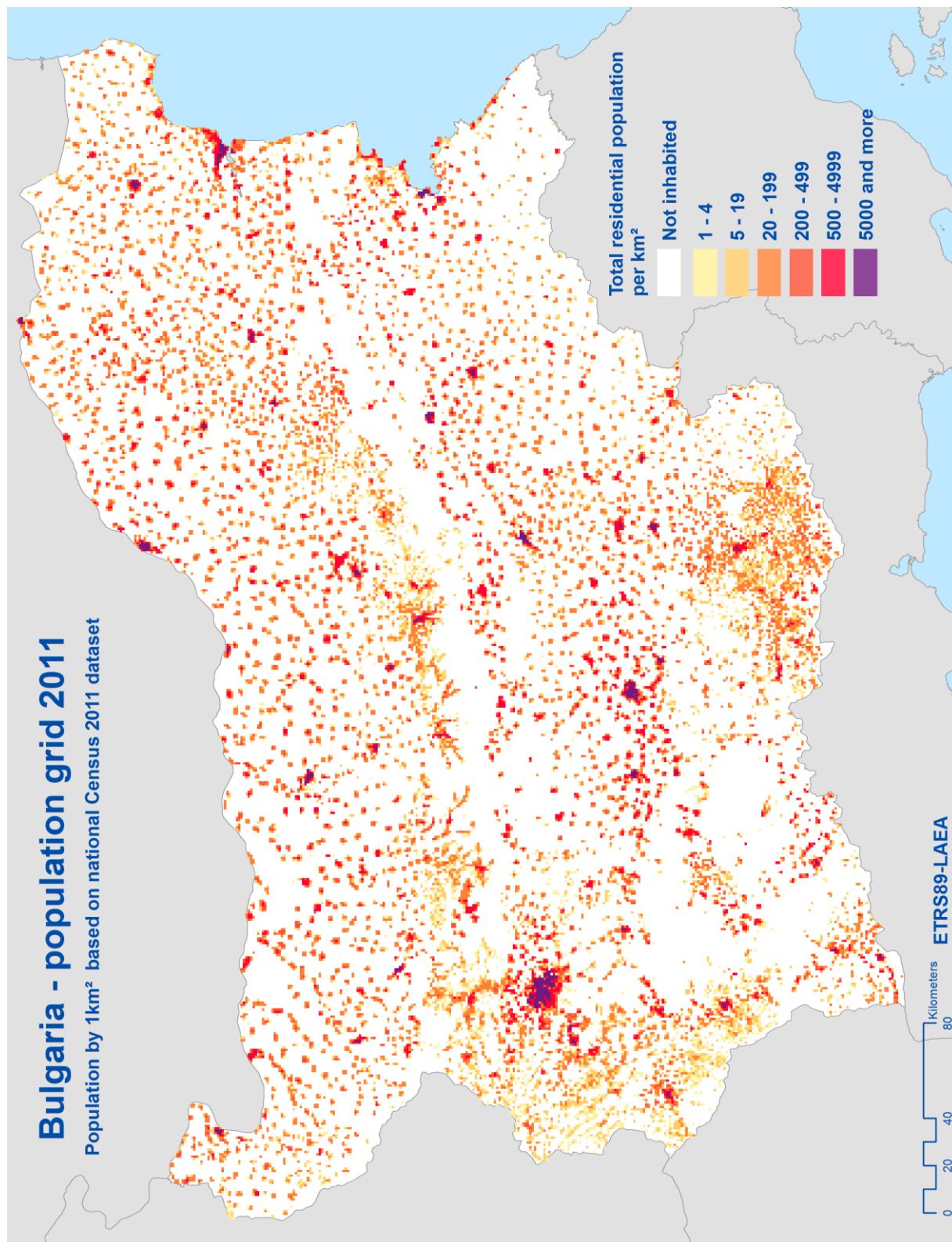
1. **EFGS**, (2011). “GEOSTAT 1A – Representing Census data in a European population grid, Final Report”.
2. **EFGS**, (2012). Guidelines from Geostat 1B “Production procedures for harmonized European population grid.
3. **Corcoran, D.** (2011). “A population grid for the Republic of Ireland: Making use of national databases and local geography”.
4. **Goerlich, F. and Cantarino, I.** (2011). “Downscaling Population with a High Resolution Land Cover Data Set for Spain”.
5. **Goerlich, F. and Cantarino, I.** (2013). “A population density grid for Spain.”
6. **Qui, F., Zhang, C and Zhou, Y.** (2012). “The development of an Areal Interpolation ArcGIS Extension and a Comparative Study”
7. **Lipatz, JL.** (2010). “Gridded population data by INSEE.”
8. **Lipatz, JL.** (2010). “Gridded data from the French census 2007. Aggregation without coordinates, coordinates but disaggregation. JL. Lipatz 23/11/2011.”
9. **Jablonski, R. and Wardzinska-Sharif, A.** (2010). “WP2 Geostatistics, Guidelines for producing statistics by grids - Poland”
10. **Steinnocher, K., Kaminger, I., Kostl, M. and Weichselbaum, J.** (2012). “Gridded Population – new data sets for an improved disaggregation approach”
11. **INSPIRE.** (2011). ”Data specification on Statistical Units – Draft Guidelines” D2.8.III.1”
12. **INSPIRE.** (2011). ”Data specification on Population Distribution - Demography – Draft Guidelines” D2.8.III.10”

Appendix I: Process for generation population grid

*UCLU – unified classifier of localization units

*UCPP – unified classifier of populated places



Appendix II: Population grid 2011 of Bulgaria

Appendix III: Population grid 2011 of Bulgaria by method