GEOSTAT 1B FINAL REPORT

Version history			
V X.X	**YYYY-MM-DD**	**Creator(s)**	**Comments**
V 1.0	2013-12-30	Oln	
V 2.0	2014-01-17	Ekp	
V 3.0	2014-02-05	Oln	
V 4.0	2014-02-06	Frb	
Project	ESSnet project GEOSTAT 1B - Representing Census data in a European population grid		
Agreement Number	50502.2009.004-2011.536		
WP	WP0 Management		
WP leader	Ola Nordbeck (Statistics Norway)		
Task			
Deliverable	Final Report for 2012-2013		
Date	2013-12-30		
Contributors	National Statistical Institute of Bulgaria	Arslan Ahmedov, Irena Dudova	
	Czech Statistical Office	Jaroslav Kraus, Štepán Moravec	
	Statistics Estonia	Ülle Valgma, Kreet Marsik	
	Statistics Finland	Marja Tammilehto-Luode, Rina Tammisto,	
	Statistics Norway	Vilni Verner Holst Bloch	
	Statistics Portugal	Ana Maria Santos	
	MD Mapping	Lars H. Backer	

ABSTRACT

This report summarizes the efforts made in the GEOSTAT 1B project. The report follows the structure of the production chain for statistics in a statistical office to underline the GEOSTAT1B project partners' emphasis on the importance of integrating the geographical component into the statistical production system. The report will bring you as a reader from the first phase of user needs analysis to the final phase of dissemination. The project has along the action generated a series of guidelines and training material beneficial to the whole European Statistical System (ESS). The deliverables discussed by the whole of the project consortium are placed in appendices, while individual contributions and other relevant material are referred to with hyperlinks.

Oslo Monday, March 10, 2014

TABLE OF CONTENTS

1. Introduction 4

2. Population grid statistics as part of the statistical production chain 5

   The Statistical production chain 5

   2.1. Specify needs 6

   2.2. Design 9

   2.3. Build, Collect and Process 13

   2.4. Analyse 16

   2.5. Disseminate 21

3. Overarching processes – Quality Management 21

   3.1 Improving the production system 21

   3.2 Improving data quality 21

6. Appendices 22

7. References 23

1. Introduction

The GEOSTAT 1B action was about developing guidelines for datasets and methods to link population and housing census 2011 statistics to a common harmonized grid. This action was launched at the beginning of 2011 and lasted for 24 months.

GEOSTAT 1B was built on the network and work made by partners in the European Forum for Geostatistics (EFGS), and the work made in the GEOSTAT 1A project[1].

Based on a series of workshops throughout the project the production of the GEOSTAT 1B material took form. This form of iterations enabled the project partners to get a better understanding of differences and challenges they were facing when working with Geographical Information Systems (GIS) in their Statistical Institutes. The material was continuously uploaded on the website (www.EFGS.info) throughout the action. Non-participating National Statistical Institutes were informed by the updates on the website and at the joint National Statistical Institute (NSI) and National Mapping and Cadastral Agency (NMCA) working group meeting organised by Eurostat and at the EFGS Conferences.

This action has received many data contributions from various partners proving that the material developed and posted on the EFGS.info website actually has been used.

Throughout the action we have tried to merge the production chain of GIS with the corresponding production chain for statistical institutes. In this report we have therefore used the structure of a statistical production chain. The deliverables discussed by the whole of the project consortium are placed in appendices, while individual contributions and other relevant material are referred to with hyperlinks.

[1] GEOSTAT 1A Consortium, 2012, GEOSTAT 1A – Representing Census data in a European population grid – Final technical report.

2. Population grid statistics as part of the statistical production chain

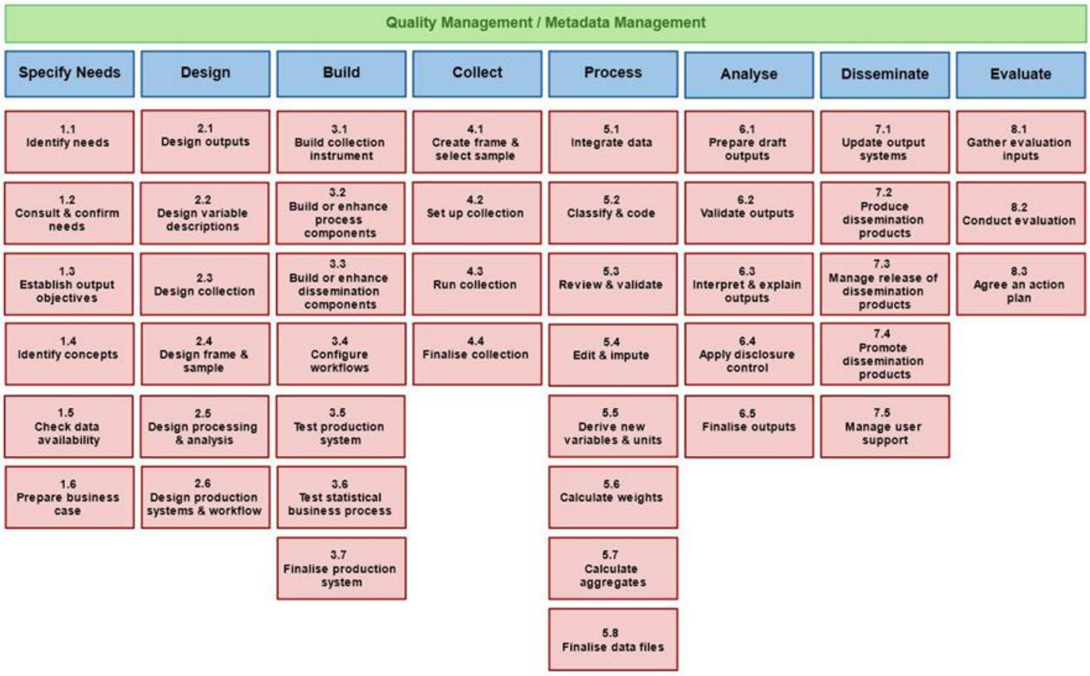
Throughout the GEOSTAT 1B project the focus has been on increasing the use of GIS (Geographical Information Systems) in the work of the National Statistical Institutes (NSI). In order to see how GIS ideally could fit into the work of a NSI, the project decided to make use of the Statistical production chain of a NSI to identify relations and enable coherence.

The Statistical production chain

The "Statistical production chain" is intended to provide a basis for assessing the possibilities and the need for standardisation and improvement of processes within NSIs. The model is also known as the Generic Statistical Business Process Model (GSBPM) and has a widespread support in the international statistical community. This is a tool in planning new statistics, in activities to improve existing work processes in statistical production, and for training purposes. It is estimated that the positive impacts of using the model primarily pertain to reduction of risk, better documentation of the production processes, e.g. through common terminology, and easier training, integration and rotation of staff.

The model describes the business processes in the statistical production on three levels, with increasing level of detail. Level 1 provides a general description of the main elements in establishing new statistics, from when the need for the statistics arises until the statistical product has been disseminated to the users[2].

The presentation of the results of the GEOSTAT 1B project in this report is based on a selected structure as proposed by the GSBPM. This report can be seen as the continuation of a presentation held by the National Statistical Institute of Bulgaria and Statistics Norway at the 2012 EFGS Conference raising the question: "An ideal way of integrating the spatial dimension into the statistical production chain - Does it exist?"[3]



Quality Management / Metadata Management

Specify Needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Build collection instrument	4.1 Create frame & select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult & confirm needs	2.2 Design variable descriptions	3.2 Build or enhance process components	4.2 Set up collection	5.2 Classify & code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Build or enhance dissemination components	4.3 Run collection	5.3 Review & validate	6.3 Interpret & explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame & sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit & impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing & analysis	3.5 Test production system		5.5 Derive new variables & units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare business case	2.6 Design production systems & workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production system		5.7 Calculate aggregates			
				5.8 Finalise data files			

FIGURE 1. THE GENERIC STATISTICAL BUSINESS PROCESS MODEL (GSBPM) BY UNECE[4]

[2] UNECE, 2013, Generic Statistical Business Process Model – GSBPM – Version 5.

[3] Ahmedov, Arslan and Nordbeck, Ola Erik, 2012, Spatial dimension into the statistical production chain.

[4] Ibid.

2.1. Specify needs

The GEOSTAT project has identified Step 1: “Specify needs” as the crucial part in the integration process.

In the call for proposals to the GEOSTAT 1B project there were two overall user needs specified:

- The need for integrating geography and statistics
- The need for spatially referenced statistics in a hierarchical system of stable and neutral grids

Integration of geography and statistics

The need for integrating geography and statistics has been stressed in various forums throughout the GEOSTAT 1B project; at the EFGS conferences, the DGINS (Directors General of the National Statistical Institutes) meeting in Prague (2012) and in late October 2013 at the first meeting of the United Nations Expert Group on the Integration of Statistical and Geospatial Information.

In the GEOSTAT 1B call for proposals this integration was described as *“The world has to be described as a set of objects. Objects are modelled in terms of their spatial aspects (geography) and in terms of their properties (statistics). A qualified description and analysis of issues therefore requires a proper integration of geography and statistics.”*

According to the call for proposals it was primarily information related to the population and housing census activities that should be treated in this project. However, according to the conclusions of the first meeting of the UN Expert Group on the Integration of Statistical and Geospatial Information, also other themes should be handled. Examples of relevant themes mentioned in the conclusions were agriculture and economic censuses and environmental-economic accounting.

In the GEOSTAT 1B project the project partners followed up on how an integration of geography and statistics can be of value in relation to the EU 2020 strategy on sustainable and inclusive growth, and more specifically to develop case studies measuring the achievements of targets in terms of resource efficiency, ageing society, and territorial cohesion (see section 2.4.2 and appendix 11).

The need for spatially referenced statistics in a hierarchical system of stable and neutral grids

The user needs for detailed population data are widespread and these needs were described in the GEOSTAT 1A report (see final technical report[5]). However, official statistical data in the ESS are in the majority of cases reported at NUTS areas which are diverse in terms of size and population and are subject to changes over time. These administrative areas, whilst suitable for reporting purposes, are therefore not always suitable for the kind of spatial analysis needed to understand the causes and consequences of social and environmental processes.

A suitable way to establish the connection between reporting units of statistical information and spatial distribution of causes and drivers is to store spatially referenced statistics in a hierarchical system of stable and neutral grids. Such a grid system can serve as a foundation for spatial analysis across all disciplines, especially for demographical, environmental and socioeconomic descriptions of human societies and their living environment. This system should not replace the existing NUTS classification, but complement it where the NUTS have limitations.

[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco_Geographical_information_maps/documents/ESSnet%20project%20GEOSTAT1A%20final%20report_0.pdf

2.1.1 Identify, consult and confirm needs for information

The overall need for population statistics on grids is described above as well as in the final technical report of GEOSTAT 1A[6]. One requirement for the GEOSTAT 1B project was to demonstrate the usefulness of grid statistics with concrete case studies and in this way raise the National Statistical Institutes' interest in producing population grids. The project identified two main types of needs:

1. The need from users for grid data as a tool for analysis and information.
2. The need from producers of statistics to produce grid data or to support the production of statistics using grid data, e.g. in the sample design.

The producer needs will be discussed in more detail in the section 3 "Build, Collect and Process", and will be referred to as **production case study**. To illustrate the user needs the project consortium proposed a series of case studies where the results can be related to EU2020 targets. Another aim was to use case studies to test the grid dataset and this is described in Section 3.2 "Test statistical business process". The following topics were proposed, also called **operative case study**[7]:

1. Health care services (capacity and location) and access to these services for elderly people

2. Analysing regional differences of tertiary educational accomplishment and to define the highly educated centres and, on the contrary, under-educated peripheries

3. Studying the ageing society and changes in education based on age structure in combination with other data sources, e.g. vital statistics or social statistics

4. Combining satellite imagery of high temporal resolution with building register data for identifying changes in green urban areas from 2000 until today

These topics were presented at the Eurostat joint working group meeting for National mapping agencies (NMA) and National statistical institutes (NSI) in 2013. As a result of this consultation process, the project consortium agreed to proceed with a modified version of topic 1 and a full report is to be found in appendix 11: "Access to emergency hospitals". The members of the yearly joint NSI and NMA GISCO working party also expressed an interest in item 4, and this case study can be found in appendix 14. This case study is an example of a hybrid approach (for explanation of hybrid approach see textbox 1 below) with the aim to detect changes in vegetation in urban areas.

2.1.2 Establish output objectives

In GEOSTAT 1A the objective was to produce population grids with the total population while in GEOSTAT 1B the overall objective was to produce population data on grids including the following variables:

- Total population, population divided by age and sex.

Another objective was to explore the consequences of introducing population breakdowns in regard to confidentiality policies in various NSIs. This implies that the study considered the consequences of various confidentiality thresholds used by NSIs when publishing population data. In the case of population grids

[6] GEOSTAT 1A Consortium, 2012, GEOSTAT 1A – Representing Census data in a European population grid – Final technical report.

[7] Nordbeck, O, 2013, Results of GEOSTAT1B – An attempt to help others.

this means the minimum number of persons in each cell that can be published without having to suppress the data.

2.1.3 CHECK DATA AVAILABILITY

In the project consortium the various project partners had access to different types of statistical and geospatial data, but for the production case study almost all project partners generated population grids using an aggregation method. The National Statistical Institute of Bulgaria combined their aggregated data with disaggregated data thereby adopting a hybrid approach (see textbox 1).

Despite a number of voluntary contributors to the GEOSTAT dataset, a number of countries have not yet produced national grid data. To achieve full EU + EFTA coverage the national population data were complemented with disaggregated population data (developed by Eurostat in cooperation with the Austrian Institute for Technology (AIT).

Textbox 1. Explanation of various approaches or methods used for producing population grids

Aggregation method: Producing grids by aggregating geo-referenced micro data (also called bottom-up approach)

Disaggregation method: In the absence of geo-referenced micro data this method produces grids, using statistical data for the lowest available administrative/territorial units in combination with auxiliary spatial data (also called top-down approach)

Hybrid method: Producing grids by combining the aggregation and disaggregation method

DATA FOR AGGREGATION FROM SPATIALLY REFERENCED POINT BASED REGISTER DATA USED IN THE PROJECT

-Address points, e.g. from Cadastre, with data from population register and/or Census 2011 statistical dataset

-Building points, from Cadastre, with data from population register and/or Census 2011 statistical dataset

-Centroids of parcels, from Cadastre

The Cadastre and population data are among the project partners often gathered from other register keepers than the statistical offices. In most of the cases they have been modified by the National Statistical Institute. This is further described in the production case description of each partner in the report “Production procedures for a harmonised European population grid – Aggregation method”, page 37-44. This report is included as appendix 1 of this report.

DATA FOR DISAGGREGATION USED IN THE PROJECT

The National Statistical Institute of Bulgaria complemented their aggregated data with the following disaggregated and auxiliary data:

-Borders of administrative units – Bulgarian Ministry of Agriculture and Foods

-Borders of populated and build-up areas (Urban land cover/land use) – Bulgarian Ministry of Agriculture and Foods

-Urban Atlas - European Environment Agency (EEA)

-Localities (partially delineated urban areas) – Bulgarian National Statistical Institute

Finally the Austrian Institute of Technology used the following data sets to produce the disaggregated population grid:

- High Resolution imperviousness layer 2009 (20m resolution)

- Imperviousness Change layer between 2006 and 2009 (20m)[8]

- Census population data per LAU2 (LAU1) 2011

- LAU2 areas 2011

- Corine Land Cover 2006[9]

- Open Street Map data[10]

OTHER DATA

Once the project partners had assembled and geo-referenced their population data, they used a standardized 1km x 1km grid net following the INSPIRE specification for the theme Geographical grid systems[11]. This multipurpose Pan-European standard is based on the ETRS89 Lambert Azimuthal Equal Area coordinate reference system. This original Pan-European grid net has been clipped per country by Eurostat in order to make the dataset lighter and easier to handle when populating the dataset. The "country clips" can be found at: http://www.efgs.info/data/eurogrid.

2.2. DESIGN

2.2.1 DESIGN OUTPUTS

THE GRID DATASET

Based on the experiences from GEOSTAT 1A, Eurostat has developed data descriptions to be used in the GEOSTAT 1B project. The project partners collected and disseminated data based on the European LAEA grid. The actual grid dataset consists of a semicolon delimited csv file with the unique INSPIRE grid cell code as reference to the grid net. The file naming of datasets followed the following standard:

Name of the file	GEOSTAT_grid_POP_1K_CC_YYYY (CC: country code, YYYY: ref. year, e.g. GEOSTAT_grid_POP_1K_SE_2011

In order to integrate all "country clips" into a European dataset a standardized table structure with pre-defined primary keys (marked with *) was necessary:

Column Names	GRD_ID*; METHD_CL; YEAR; CNTR_CODE*; DATA_SRC; TOT_P; TOT_F; TOT_M; F_00_14; F_15_64; F_65_; M_00_14; M_15_64; M_65_

[8] EEA, 2009, Technical Note on HR Imperviousness Layer Product Specification.

[9] EEA, 2009, *Corine Land Cover 2006.*

[10] http://www.openstreetmap.org/.

[11] http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_GGS_v3.0.1.pdf.

Where:

GRD_ID	Identification code of the grid cell (based on the LAEA grid) starts with the grid cell size (e.g. 1km) followed by the coordinates of the lower left-hand corner. The coordinates are in km and start with the letter "N" followed by the latitude and "E" followed by the longitude This results in the following GRD_ID: 1kmN4101E4453
METHD_CL	Method used to determine the population of the grid cell; A (aggregated), D (disaggregated) and M (mixed). In the case of border cells calculations had to be made in order to remove double counting of the population.
YEAR	Reference year of the data
CNTR_CODE	ISO code of the country in which the grid cell is located (in the case of border cells the grid cell is reported from all neighboring countries with the same GRD_ID, but different CNTR_CODE attribute).
DATA_SRC	For national datasets the country code; for the European disaggregated datasets the data source. For border cells a similar approach as for CNTR_CODE was adopted.

THE POPULATION VARIABLES

Making use of the grid dataset the partners agreed to recommend producing a range of variables as official statistics on 1km² grids. This means that this information would be free of charge. These recommendations can be found in the document "EFGS Standard for official statistics, population variables V.1.0" (see appendix 2). These population data variables are all linked to the population at their usual place of residence as laid down in Article 2(d) of Regulation (EC) No 763/2008.

For the GEOSTAT 1B population grid the following census topics were requested:

TOTAL POPULATION

This is the simplest way of displaying population statistics, with only total population per grid cell.

Name of the variable	Definition
TOT_P	[Number of inhabitants, total]

POPULATION BY SEX, TOTAL

In addition to total population per grid cell population can also be divided by sex.

Name of the variable	Definition
TOT_M	[Number of inhabitants, male]
TOT_F	[Number of inhabitants, female]

POPULATION BY SEX AND 3 AGE GROUPS.

This is the most detailed population grid statistics recommended by EFGS as official grid statistics.

Legal documents (EU) relevant to this grouping: Council Regulation (EC) nos 577/98, 1991/2002 and 2257/2003, and Commission Regulation nos 1575/2000, 1897/2000, 2104/2002, 430/2005 and 377/2008.

Name of the variable	Definition
M_00_14	[Number of inhabitants, male age 0-14]
M_15_64	[Number of inhabitants, male age 15-64]
M_65_	[Number of inhabitants, male age 65-]
F_00_14	[Number of inhabitants, female age 0-14]
F_15_64	[Number of inhabitants, female age 15-64]
F_65_	[Number of inhabitants, female age 65-]

DISCLOSURE CONTROL

In line with the recommendations of GEOSTAT 1A the project partners in GEOSTAT 1B agreed that the total population should be considered as official statistics; free of charge and without disclosure to be reported on 1 km^2 grid cells.

In case data disclosure is considered, the project initially defined 10 as an appropriate all-round threshold, but as described in chart 2 (se page 20) this value is too high to be applied in case of multiple population break-downs (variables). Statistics Finland practiced here a valid alternative that was based on the disclosure risk of the population breakdowns per age or sex for all 1 km^2 grid cells. If the total population was under 10 persons, than the population breakdowns were protected, but not the total population.

2.2.2 DESIGN OF METADATA AND QUALITY ASSESSMENT PARAMETERS

The following quality assessment parameters are based on the initial work carried out in the GEOSTAT 1A project. The GEOSTAT 1B action discussed these parameters in a first iteration in 2012. Based on feedbacks in 2013 from various European NSIs outside the GEOSTAT 1B action, these parameters were further developed and finalised by late 2013. The parameters cover a form of combined metadata description and quality assessment of the primary data as well as quality assessment of the grid data production. The data are to be filled in as a word document following the structure below.

The file naming of this word document should follow the following standard:

Name of the file	GEOSTAT_grid_POP_1K_CC_YYYY_QA (CC: country code, YYYY: ref. year, QA: quality assessment, e.g. GEOSTAT_grid_POP_1K_SE_2011_QA

PARAMETERS FOR MEASURING THE QUALITY OF THE PRIMARY DATA PRODUCTION

Type of primary data	Geo-referenced data in the form of point, line or polygon
Positional accuracy	Accuracy of geo-referenced data in meters for points or by scale for lines and polygons.
Positional source	Are the geo-referenced data the result of a computation from official data sources or are they interpolated between known points (e.g. between addresses at cross roads). The 'approximately located population proportion' of the INSPIRE specification provides information about the consistency of geo-reference measurements.
Logical consistency	Yes/no (no= different types of geo-referenced data)
Usage	How can the data be used and what are the limitations. E.g. population registers are based upon where people have their primary dwelling, and hence geocodes people to where they mostly stay at night. However, part of the population have several dwellings; for recreation, work, shared custody etc. This implies that there are some constraints to which analyses may and can be done. Furthermore, a large share of the population may work at night or abroad. This further limits the use with respect to daytime phenomena.
Bias	Individual data that are recorded and counted in some places, but only on the basis of some convention: e.g. homeless people in the place where the data are collected during a standard census process or in the place of the organisation that takes care of them for social benefits or health insurance purposes. This is covered by the 'conventionally located population proportion' of the INSPIRE specification.
Accuracy of the figures	In the cases of figures produced by extrapolation of the sample (mainly censuses) the measure of accuracy that results from the estimation process. Percentage of non-matching points.
Coverage of geo-referenced data	% of statistical (primary) data covered with geo-referenced data. This is covered by the 'not counted population proportion' of the INSPIRE specification.
Date of latest update of the spatial data and percentage updated	Last updates of the geo-referenced data or the territorial units used for disaggregation. Extent of the updated data (%).
Date of latest update of the statistical data and percentage updated	Reference day or last update of the data itself (e.g. census decisive moment).This is covered by the 'period of measurement' of the INSPIRE specification that can be delivered for the whole data set and optionally for each data cell if there are exceptions.
Coherence	% of consistent and comparable data — regional differences in quality. If possible.
Quality report	Available / not available (If available, url or reference to quality report should be given.)
Inspire compliant metadata	Available / not available

PARAMETERS FOR MEASURING THE QUALITY OF THE GRID DATA PRODUCTION

Production methods	Aggregation, disaggregation, mixed mode. The aggregation method may have subgroups: aggregation direct from register data, aggregation made on estimated individual data based on registers, and aggregation from very small enumeration areas. Disaggregation methods may have subgroups depending on the regional level of the source data and the ancillary data used.
Accuracy of the figures	In the cases where figures are produced by extrapolation of the sample (mainly censuses) the measure of accuracy that results from the estimation process. Difference between the population number in grids and the reported total population.

Geographical coverage	Percentage of country area covered by grid data
Coherence	% of consistent and comparable data — regional differences in quality. If possible.
Temporal accuracy	% of same reference date.
Confidentiality	% of population* suppressed, % of grid cells suppressed, thresholds for confidential data. (population* can also be other statistical variables as dwellings etc.)
Inspire compliant metadata	Available / not available
License	Available / not available (If available, an abstract, an url or reference to license agreement should be given.)

Section 1.3 “Check data availability” provides an overview of the data used by the different GEOSTAT 1B partners.

A detailed discussion of data sources used can be found on page 37-44 in appendix 1: “Production procedures for a harmonised European population grid – Aggregation method”.

The population data contributions from various European NSIs include also the quality assessment sheets. These sheets also provide a better understanding of differences in the final output: the population grid data.

In the document “Testing and quality assessment of Pan-European grids” in appendix 3 there is an overview of the disaggregated data used in the disaggregation work of Austrian Institute of Technology. This document also gives a brief description of how national ancillary data can be used in order to improve the quality of the disaggregation.

2.3. Build, Collect and Process

Based on the design of the methodological framework the various partners carried out their work with building, collecting and processing their data. This work resulted in short production case studies per project partner (see appendix 1, “Production procedures for a harmonised European population grid – Aggregation method”). In the iterative process of the action these production case studies were used for adjusting the methodological framework.

Within the GEOSTAT 1B action various NSIs had access to different types of primary data. Depending on the resolution and accuracy of these primary data, this resulted in two different approaches to producing population grids: “Aggregation” and “Hybrid”.

The aggregation design was developed in an iterative process and was initially based on Statistics Finland’s approach; see figure 2and appendix 4: “Production process Bottom-Up Finland”.

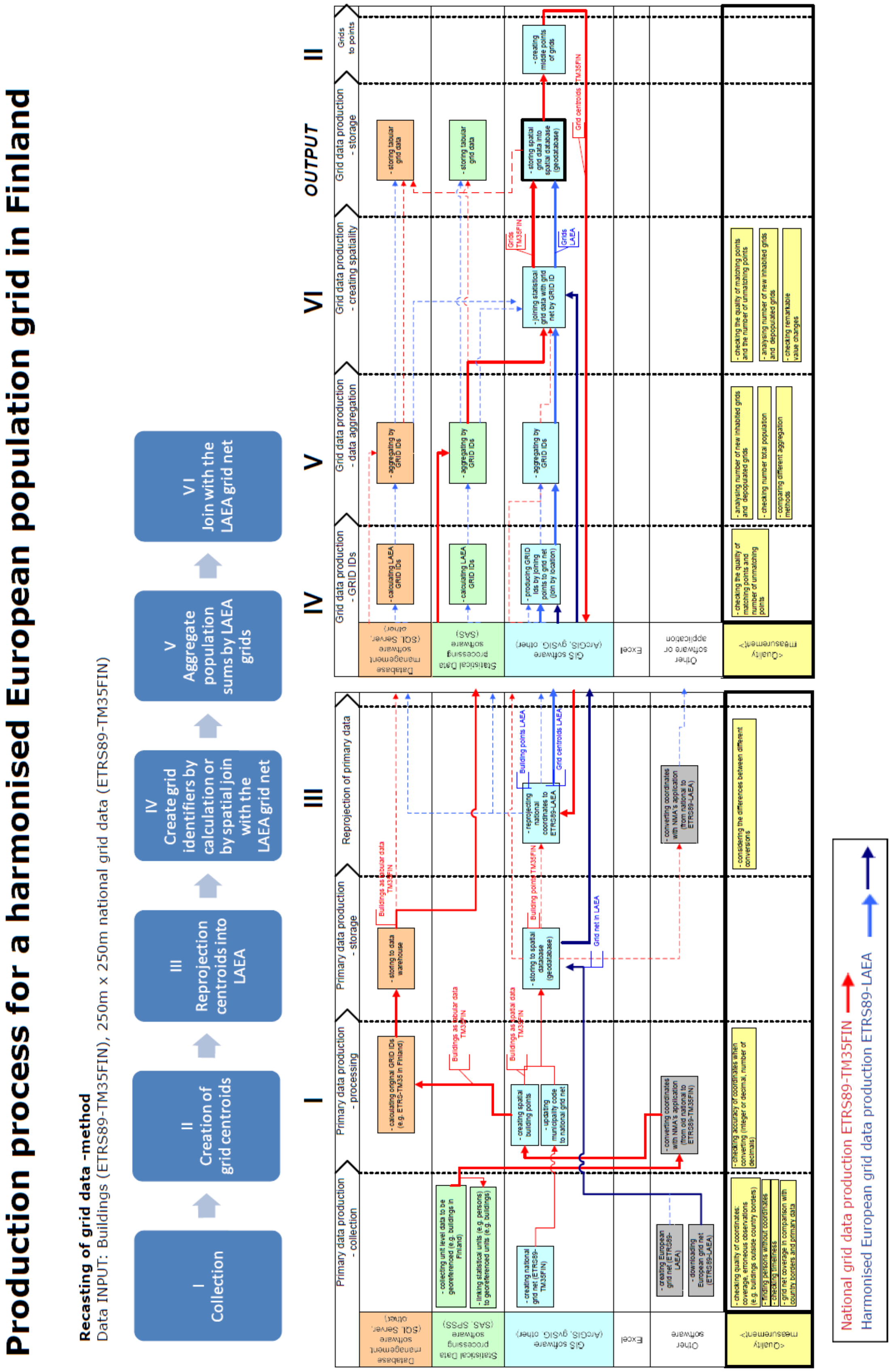
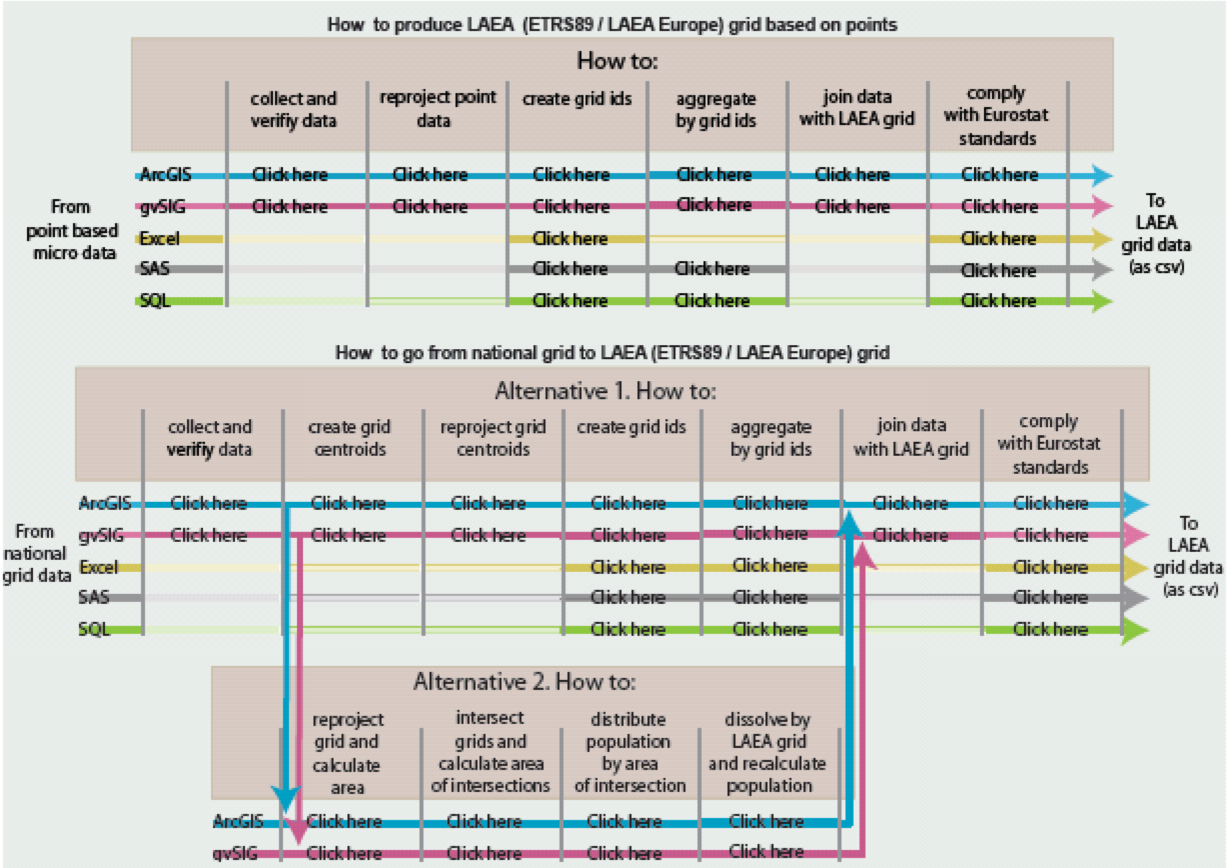

Production process for a harmonised European population grid in Finland
Recasting of grid data -method
Data INPUT: Buildings (ETRS89-TM35FIN), 250m x 250m national grid data (ETRS89-TM35FIN)
I Collection
II Creation of grid centroids
III Reprojection centroids into LAEA
IV Create grid identifiers by calculation or by spatial join with the LAEA grid net
V Aggregate population sums by LAEA grids
VI Join with the LAEA grid net
I
III
IV
V
VI
OUTPUT
II
Primary data production - collection
Primary data production - processing
Primary data production - storage
Reprojection of primary data
Grid data production - GRID IDs
Grid data production - data aggregation
Grid data production - creating spatiality
Grid data production - storage
Grids to points
Database management software (SQL Server, other)
Statistical Data processing software (SAS, SPSS)
GIS software (ArcGIS, gvSIG, other)
Excel
Other software
Other software or application
<Quality measurement>
- collecting unit level data to be georeferenced (e.g. buildings in Finland)
- linking statistical units (e.g. persons) to georeferenced units (e.g. buildings)
- creating national grid net (ETRS89-TM35FIN)
- creating European grid net (ETRS89-LAEA)
- downloading European grid net (ETRS89-LAEA)
- calculating original GRID IDs (e.g. ETRS-TM35 in Finland)
- creating spatial building points
- upgrading municipality code to national grid net
- converting coordinates with NMA's application (from old national to ETRS89-TM35FIN)
- storing to data warehouse
- storing to spatial database (geodatabase)
- reprojecting national coordinates to ETRS89-LAEA
- converting coordinates with NMA's application (from national to ETRS89-LAEA)
Buildings as tabular data TM35FIN
Buildings as spatial data TM35FIN
Building points TM35FIN
Building points LAEA
Grid centroids LAEA
Grid net in LAEA
- calculating LAEA GRID IDs
- producing GRID Ids by joining points to grid net (join by location)
- aggregating by GRID IDs
- joining statistical grid data with grid net by GRID ID
Grids TM35FIN
Grids LAEA
- storing tabular grid data
- storing spatial grid data into spatial database (geodatabase)
Grid centroids TM35FIN
- creating middle points of grids
- checking quality of coordinates: coverage, erroneous observations (e.g. buildings outside country borders)
- finding persons without coordinates
- checking timeliness
- grid net coverage in comparison with country borders and primary data
- checking accuracy of coordinates when converting (integer or decimal, number of decimals)
- considering the differences between different conversions
- checking the quality of matching points and number of unmatching points
- analysing number of new inhabited grids and depopulated grids
- checking number total population
- comparing different aggregation methods
- checking the quality of matching points and the number of unmatching points
- analysing number of new inhabited grids and depopulated grids
- checking remarkable value changes
National grid data production ETRS89-TM35FIN
Harmonised European grid data production ETRS89-LAEA

FIGURE 2. POPULATION GRID PRODUCTION FOLLOWING THE BOTTOM-UP METHODOLOGY, BY STATISTICS FINLAND

This approach was improved throughout the project and made more generic based on the various partners' approaches and primary data. This resulted in a methodology paper: "Production procedures for a harmonised European population grid – Aggregation method" (appendix 1), a poster: "Contribute to the GEOSTAT1B population grid" (appendix 5) and a web tool (http://www.efgs.info/geostat/1B/training-material) for making the population grid production easier (see figure 3).

These products were produced by the GEOSTAT project in order to assist National Statistical institutes in production of population data on grids. The methodology report is based on the experiences from the various partners. The training material gives the user various alternative approaches for generating grids using ArcGIS- or Open Source GIS software, Statistical software as SAS/Excel or using Database queries (SQLs).



How to produce LAEA (ETRS89 / LAEA Europe) grid based on points

How to:

From point based micro data	collect and verify data	reproject point data	create grid ids	aggregate by grid ids	join data with LAEA grid	comply with Eurostat standards	To LAEA grid data (as csv)
ArcGIS	Click here	Click here	Click here	Click here	Click here	Click here	
gvSIG	Click here	Click here	Click here	Click here	Click here	Click here	
Excel			Click here			Click here	
SAS			Click here	Click here		Click here	
SQL			Click here	Click here		Click here	

How to go from national grid to LAEA (ETRS89 / LAEA Europe) grid

Alternative 1. How to:

From national grid data	collect and verify data	create grid centroids	reproject grid centroids	create grid ids	aggregate by grid ids	join data with LAEA grid	comply with Eurostat standards	To LAEA grid data (as csv)
ArcGIS	Click here	Click here	Click here	Click here	Click here	Click here	Click here	
gvSIG	Click here	Click here	Click here	Click here	Click here	Click here	Click here	
Excel				Click here	Click here		Click here	
SAS				Click here	Click here		Click here	
SQL				Click here	Click here		Click here	

Alternative 2. How to:

	reproject grid and calculate area	intersect grids and calculate area of intersections	distribute population by area of intersection	dissolve by LAEA grid and recalculate population
ArcGIS	Click here	Click here	Click here	Click here
gvSIG	Click here	Click here	Click here	Click here

FIGURE 3. WEB TOOL FOR POPULATION GRID PRODUCTION FOLLOWING THE BOTTOM-UP METHODOLOGY

This training material for the aggregation approach was further developed by the National Statistical Institute of Bulgaria in their population grid production using the hybrid approach. Figure 3 and appendix 6 describe three steps where the first one refers to the aggregation method while step two and three refer to the disaggregation approach (see section 1.3 "Identify data sources and data availability"). This approach is further described in appendix 9.

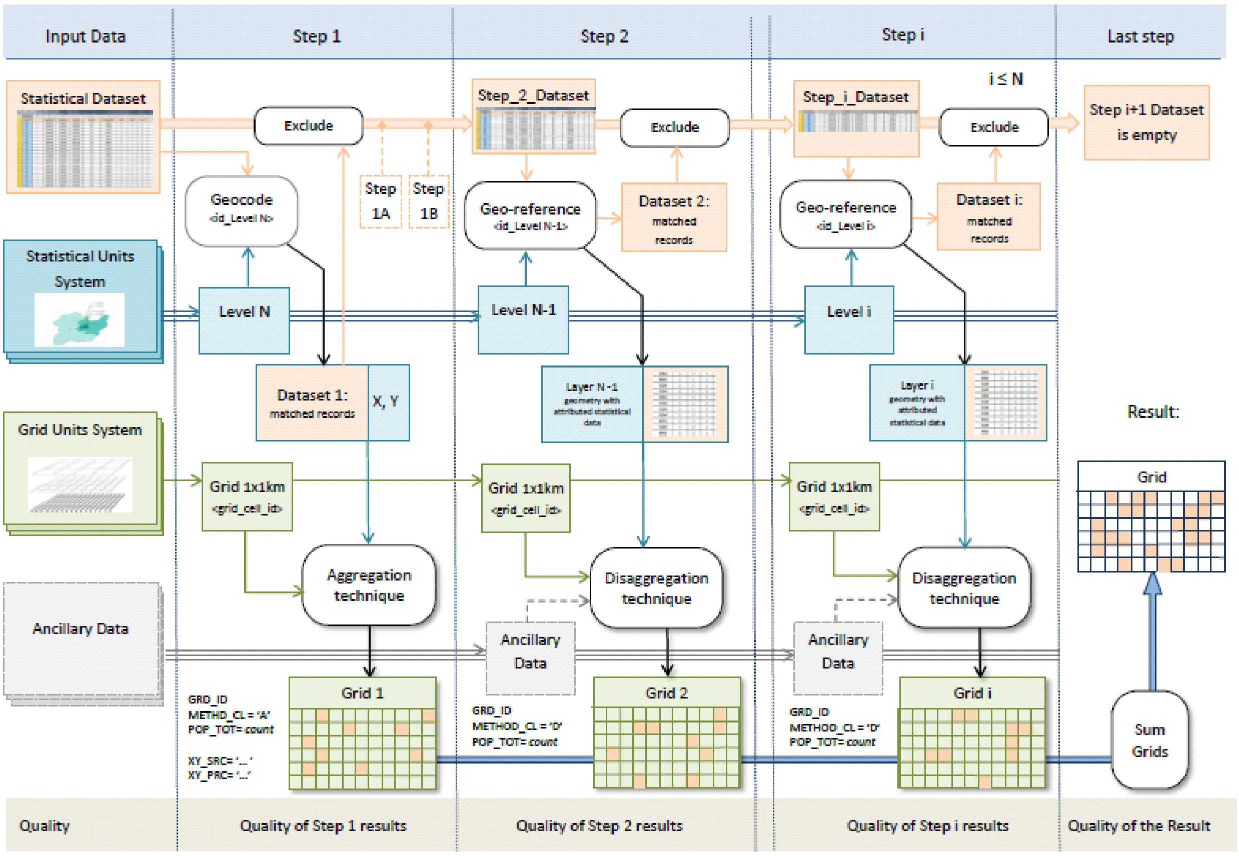


FIGURE 4. POPULATION GRID PRODUCTION FOLLOWING THE HYBRID METHODOLOGY OF NATIONAL STATISTICAL INSTITUTE OF BULGARIA

Based on the design of the methodological framework the various partners carried out their work with building, collecting and processing their data. This work resulted in short production case studies per project partner (see appendix 1, “Production procedures for a harmonised European population grid – Aggregation method”). In the iterative process of the action these production case studies were used for adjusting the methodological framework.

Based on these amendments and the updates of the training material an invitation letter was sent out in March 2013 asking all European NSI (EU27 +) to use the material and to contribute to GEOSTAT population grid.

The hybrid approach used by the National Statistical Institute of Bulgaria was presented at the EFGS conference in Sofia 2013 (see appendix 8, “Prototype of Bulgarian population grid 2011”) and the approach is further described in this production case report: (see appendix 9, “Summary report of the methodology for generating Bulgarian population grid 2011”).

2.4. ANALYSE

2.4.1 PREPARE DRAFT RESULTS

As a response to the data request from the project, the following countries contributed to the GEOSTAT 2011 grid:

Belgium, Bulgaria, Czech Republic, Estonia, Finland, France, Ireland, Netherlands, Norway, Poland, Portugal, Slovenia, Sweden and Switzerland.

Some NSIs are about to finalise their census data for 2011 and have expressed an interest in sharing data with Eurostat in 2014/2015.

The data contributions have all followed the structure in the GEOSTAT1B guidelines. This has made the integration and harmonisation work of individual contributions easier than it was in the GEOSTAT 1A project.

A draft poster of the GEOSTAT 2011 data Version 0.1 (October 2013)[12] was produced by Eurostat to the EFGS Conference in Sofia 2013. A version 1.0 is expected to be produced based on the contributions by January 2014.

The population data contributions from the following European NSIs: Bulgaria, Czech Republic, Estonia, Finland, France, Ireland, Netherlands, Norway, Portugal and Slovenia have all included the quality assessment forms as described under “2.2 Design of metadata and quality assessment parameters”. These forms are to be found in appendix 17 and give a better understanding of the data presented in the GEOSTAT2011 grid.

The forms in appendix 17 contain two notable different cases, the confidentiality policy in Ireland and the production method in France. Considering the differences in confidentiality thresholds the project consortium found it important to make some tests with the data shared with the project, see section 2.4.2.

2.4.2 VALIDATE OUTPUTS

In order to test the population grid, the GEOSTAT 1B partners agreed on working together on a case study calculating the population within 30 minutes from emergency hospitals. In order to calculate the travel time, GIS was used and the aim was to explore:

- the consequences of using grids as statistical units instead of other statistical units as e.g. municipality areas

- the consequences of data confidentiality policies applied by various NSIs when publishing population grids

This case study is making use of the 1km² GEOSTAT 1B population grids in combination with geo-referenced road networks and emergency hospitals in order to determine the travel time to emergency hospitals. The population within the driving distance of interest is divided into various groups of age and sex.

In order to test using grids as output areas instead of municipality areas, addresses/building points were used as a benchmark. A first test compared the effect of the selection method (centroid vs. polygon) for municipality areas and grid cells. Figure 5, below, illustrates the differences in geographical coverage. The map in the middle shows that intersecting municipality centroids with a service area results in many address/building points that are within the service area, are actually located outside the selected municipalities, and are therefore would be excluded from the journey time statistics per municipality. When intersecting with the whole municipality areas, the result is the opposite; Address/building points outside the service area fall inside the selected municipalities and are falsely included into the journey time statistics. This effect occurs only to a minor extent in the right hand map, where grid centroids and grid

[12] http://www.nsi.bg/efgs2013/data/uploads/presentations/DAY2_WS1_2_Presentation_PETRI_ok.pdf

polygons are compared which means that contrary to administrative units grid cells are robust against the intersection method

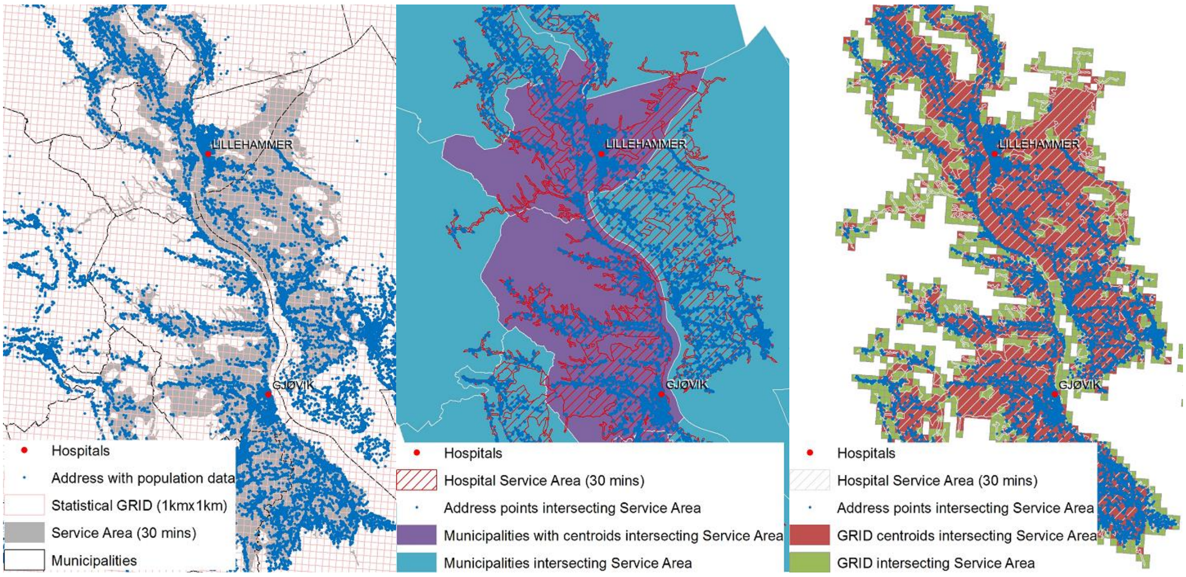

LILLEHAMMER
GJØVIK
Hospitals
Address with population data
Statistical GRID (1kmx1km)
Service Area (30 mins)
Municipalities
LILLEHAMMER
GJØVIK
Hospitals
Hospital Service Area (30 mins)
Address points intersecting Service Area
Municipalities with centroids intersecting Service Area
Municipalities intersecting Service Area
LILLEHAMMER
GJØVIK
Hospitals
Hospital Service Area (30 mins)
Address points intersecting Service Area
GRID centroids intersecting Service Area
GRID intersecting Service Area

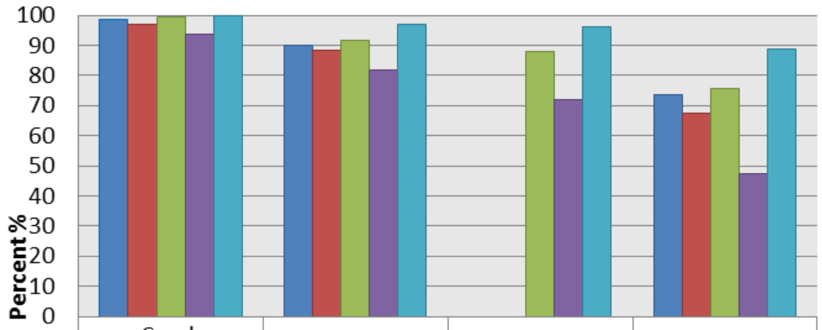
FIGURE 5: SERVICE AREAS ACCESSIBLE IN 30 MINUTES FROM EMERGENCY HOSPITALS INTERSECTING POPULATION AT DIFFERENT LEVELS OF STATISTICAL UNITS.

Left: Statistical units on various geographical levels used in the analysis

Middle: Result of intersecting the municipalities alternatively the municipality centroids with the 30 minutes service areas. Some of the municipalities have both centroids and municipality areas that intersect the service areas. In figure 5 this has as a consequence that the municipalities in violet are also included in the turquoise category.

Right: Result of intersecting the grids alternatively the centroids of the grids with 30 minutes service areas. Some of the grid cells have both centroids and grid cells that intersect the service areas. In the figure this has as a consequence that the grids in dark red are also included in the green category.

CHART 1: RESULTS OF INTERSECTING 30 MINUTE SERVICE AREAS WITH POPULATION DATA ON STATISTICAL UNITS. THE COLOR CODE IS THE SAME AS IN FIGURE 5 ABOVE.



	Czech Republic	Estonia	Finland	Norway
🟦 Intersecting address/building centroids with service areas	98,8	90,2	0,0	73,7
🟥 Intersecting 1km*1km grid centroids with service areas	96,8	88,3	0,0	67,3
🟩 Intersecting 1km*1km grid polygons with service areas	99,4	91,5	87,9	75,6
🟪 Intersecting municipality centroids with service areas	93,9	81,7	72,1	47,5
🟦 Intersecting municipality polygons with service areas	99,8	97,1	96,1	88,6

The impact of differences in geographical coverage becomes even more obvious when statistics are attached to the intersected address/building point, grids and municipalities. This is illustrated in chart 1, in which different countries are compared in terms of their population proportion living within 30 minutes service areas. The most important factor here is the size of the municipalities that varies inside and in between the countries, while the size of the grids is constant.

In chart 1 it also appears that the results for the Czech Republic are significantly better than in the other countries. The main reason for this is that the 30 minutes service areas in the Czech Republic cover most of the country’s territory while this is not the case in the other countries.

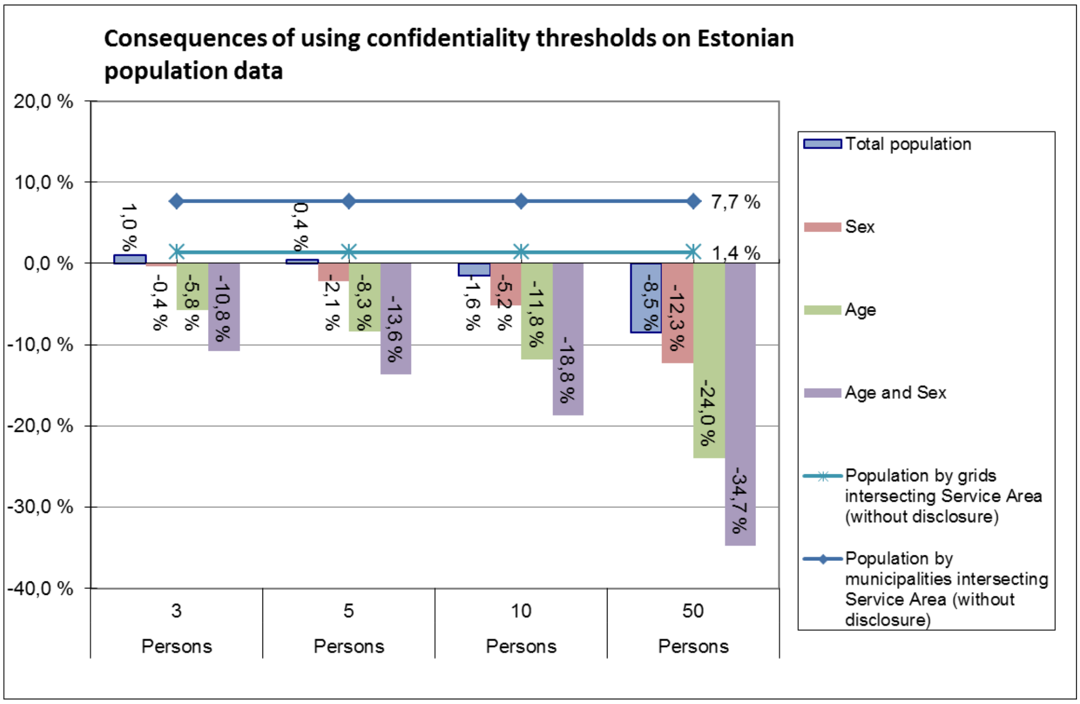
The main differences between the countries in this accessibility study lie in:

- Hospital coverage
- Population distribution
- Size and the physical geography (e.g. hilliness, coastline, lakes, islands) of the countries.
- Road network (incl. coverage and speed limits)

The results of the benchmarking in chart 1 show that the best approach is to use grid polygons. The centroids of the grids are easier to handle, but as for the municipalities the population distribution inside the grid polygon is not known. However, based on the theory that people live close to the road network the service areas might indicate where the population is situated in the grid. The case studies proved that results based on grid data are more similar to the reference of microdata associated to the building centroids than data aggregated to municipality level.

Nevertheless, introducing confidentiality thresholds and consequently suppression of grid cells below these thresholds may result in that the advantages of grid statistics are lost. This loss of grid cells and consequently population is illustrated in chart 2 where the correct population data is equal zero. By aggregating the statistics by grids or by municipalities the “overshoot” is 1,4 % and 7,7% respectively. This error margin must be compared to the “undershoot” that occurs as a consequence of data suppression in the case of multiple population breakdowns (variables) in combination with high confidentiality thresholds. This means a threshold of 5 persons and applied on a breakdown in only two classes (topic sex) results already in a protection of 2% of the population and more than 8% for three age classes.

CHART 2. BENCHMARKING POPULATION DATA FOR FINDING OUT THE CONSEQUENCES OF DATA SUPPRESSION, INSIDE 30 MINUTES SERVICE AREAS IN ESTONIA. 0 % REFER TO THE POPULATION BY BUILDING CENTROIDS.



There are more examples in the operative case study called “Access to emergency hospitals” (see appendix 11) and this approach can allow NSIs to define the detail of variables and the threshold levels. If too high thresholds are applied then reporting from a NSI might be more correction the basis of administrative areas instead of grids. In the case of administrative areas NSIs do not have to consider the confidentiality thresholds. As mentioned under section 2.2.1 Design outputs the disclosure control of Statistics Finland is a valid alternative for NSI who are interested in presenting population breakdowns, but in line with domestic confidentiality policy.

2.4.3 Finalise outputs

Efforts have been made throughout the project to make it easy to sew together all national contributions into one harmonised grid. In case the design outputs recommendations under the design phase above are followed, it will facilitate this work. In a new round of GEOSTAT it is important to develop tools for quality assessment (QA) reporting and replace the current forms created in Microsoft Words. This new tool should make it easier to structure and compare contributions from various NSI’s.

GEOSTAT 1B also followed up on previous discussions in the GEOSTAT ESSnet 1A project concerning a shared business model. In this regard the project discussed and proposed a common open licence model for European grid data sets. This was also the conclusion of a web survey carried out in 2012 (see appendix 12), where core datasets for grid statistics were suggested to be official grid statistics. For these core data sets for grid statistics, a common open data licence template (see appendix 13) has been elaborated based upon the open data licence made by the National Land Survey of Finland.

2.5. Disseminate

Several efforts have been made to advocate for the production and the use of population statistics on grids. Organising four workshops and two Conferences made it possible to have foras where it was possible to promote and exchange ideas regarding grid statistcs. The project used the EFGS website for publishing data and reports throughout the project and this website will also in the future be an important tool for dissemination. The website has been uploaded with serveral papers, presentations and reports thoughout the project.

The website will also in the future be an important tool for publishing the future GEOSTAT 2011 datasets.

3. Overarching processes – Quality Management

3.1 Improving the production system

The various GEOSTAT 1B partners have or are in the process of integrating the production of grids statistics into their regular production systems for official statistics.

Statistics Portugal also used the results of the 2011 Population and Housing Census for the constitution of a new sampling frame for household survey purposes (national buildings and dwellings register). Important aspects for this frame are the availability of the geo-referenced buildings of the 2011 census and the access to data from different administrative sources. Unlike the former master sample, this new sampling frame will not have a fixed lifetime of a decade, but will be subject to continuous updating by means of the integration of the information from the administrative sources.

The European GRID is, together with the geographical location of buildings, one of the important geographical components of the sampling frame. The use of this GRID allows the incorporation of a degree of flexibility within the sampling process, creating independence on the administrative division. For each building their corresponding GRID Cell is known. For the sampling process of the household surveys, cells are selected within strata (NUTS3 areas), and a spatial sorting algorithm assures the contiguity/proximity of each cell. The cells were also classified in High and Low Density, a necessity for some surveys; this classification is based on the population density of each cell and its adjacent cells. For more information see appendix 10: Using the European Grid “ETRS89/LAEA_PT_1K”as the foundation for the new Portuguese Sampling Infrastructure.

3.2 Improving data quality

Another paper was presented by the Czech Statistical Office and described how to handle the population that is not linked to the exact place of usual residence in the Population and Housing Census 2011.

When processing Population and Housing Census 2011, an overwhelming majority of population data was geo-referenced into building points. This makes it possible to apply the aggregation method for producing a population grid. However, there are about 93 000 persons (i.e. about 0.9 % of the total census population), who can be linked down only to the level of statistical districts, but not to the exact place of usual residence

(e.g. homeless people, people living in buildings without final approval or in emergency buildings or shelters). That means that distribution of these persons into the exact place (with x, y coordinates) or alternatively into grids must be conducted through some disaggregation method.
This test of the disaggregation method was conducted in the city of Abertamy in the northern part of the Czech Republic, very close to Germany. The paper discusses this method and indicates its advantages and disadvantages. The final implementation of the disaggregation approach for the Czech Republic will be adopted before publishing census results in grid format. For more information see appendix 16.

6. Appendices

Appendix 1. Production procedures for a harmonised European population grid - Aggregation method:

http://cros-portal.eu/content/appendix1-production-procedures-aggregation-method

Appendix 2. EFGS standard for official grid statistics, population variables v.1.0:

http://cros-portal.eu/content/appendix2-efgs-standard-official-statistics-population-variables

Appendix 3. Testing and quality assessment of pan-European population grids - based on the Norwegian experience

http://cros-portal.eu/content/appendix3-testing-and-quality-assessment-pan-european-grids

Appendix 4. Production process for a harmonised European population grid in Finland

http://cros-portal.eu/content/appendix4-poster-production-process-bottom-approach-finland

Appendix 5. Contribute to the GEOSTAT population grid 2011

http://cros-portal.eu/content/appendix5-poster-how-contribute-geostat1b-popgrid-poster-a0

Appendix 6. Production process for a hybrid approach to harmonised European population grid, Bulgaria

http://cros-portal.eu/content/appendix6-poster-production-process-hybrid-bulgaria

Appendix 8. Prototype of Bulgarian population grid 2011

http://cros-portal.eu/content/appendix8-presentation-hybrid-approach-statistics-bulgaria

Appendix 9. Methodology for generating Bulgarian population grid 2011

http://cros-portal.eu/content/appendix9-methodology-hybrid-approach-statistics-bulgaria

Appendix 10. Using the European Grid “ETRS89/LAEA_PT_1K”as the foundation for the new Portuguese Sampling Infrastructure

http://cros-portal.eu/content/appendix-10-using-european-grid-%E2%80%9Cetrs89laeapt1k%E2%80%9D-foundation-new-portuguese-sampling

Appendix 11. Access to emergency hospitals - an operative case study for testing gridded population data

http://cros-portal.eu/content/appendix-11-access-emergency-hospitals

Appendix 12. EFGS 2012 web survey

http://cros-portal.eu/content/appendix12-efgs-2012-web-survey

Appendix 13. EFGS open data license template

http://cros-portal.eu/content/appendix13-efgs-open-data-license-template

Appendix 14. Vegetation change detection methodology for statistical institutes - A case study for combining building register data with satellite imagery

http://cros-portal.eu/content/appendix-14-vegatation-change-detection-statistical-institutes-case-study-combining-building

Appendix 15. GEOSTAT 1B Flyer

http://cros-portal.eu/content/appendix15-wp4-flyer-efgs-geostat

Appendix 16. Disaggregation methods for georeferencing inhabitants with unknown place of residence: the case study of population census 2011 in the Czech Republic

http://cros-portal.eu/content/appendix-16-disaggregation-methods-georeferencing-inhabitants-unknown-place-residence-case

Appendix 17. EFGS modified quality assessment parameters, filled in by GEOSTAT1B contributors

http://cros-portal.eu/content/appendix-17-efgs-modified-quality-assessment-parameters-filled-geostat1b-contributors

7. References

Ahmedov, Arslan and Nordbeck, Ola Erik, 2012, Spatial dimension into the statistical production chain http://www.efgs.info/geostat/workshops/efgs-2012-prague-czech-republic/efgs-2012-conference/ahmedov-presentation

EEA, 2009, Corine Land Cover 2006
http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-1

EEA, 2009, Technical Note on HR Imperviousness Layer Product Specification
http://www.gmes-geoland.info/fileadmin/geoland2/redakteur/pdf/Project_Documentation/Service_Specification/TechnicalProductSpecification_HR_Imperviousness_Layer_I1-01.pdf

GEOSTAT 1A Consortium, 2012, GEOSTAT 1A – Representing Census data in a European population grid – Final technical report
http://epp.eurostat.ec.europa.eu/portal/page/portal/gisco_Geographical_information_maps/documents/ESSnet%20project%20GEOSTAT1A%20final%20report_0.pdf

UNECE, 2013, Generic Statistical Business Process Model – GSBPM – Version 5
http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

Nordbeck, O, 2013, Results of GEOSTAT1B – An attempt to help others
http://www.efgs.info/geostat/1B/frontpage/geostat-1b-presented-at-gisco-2013