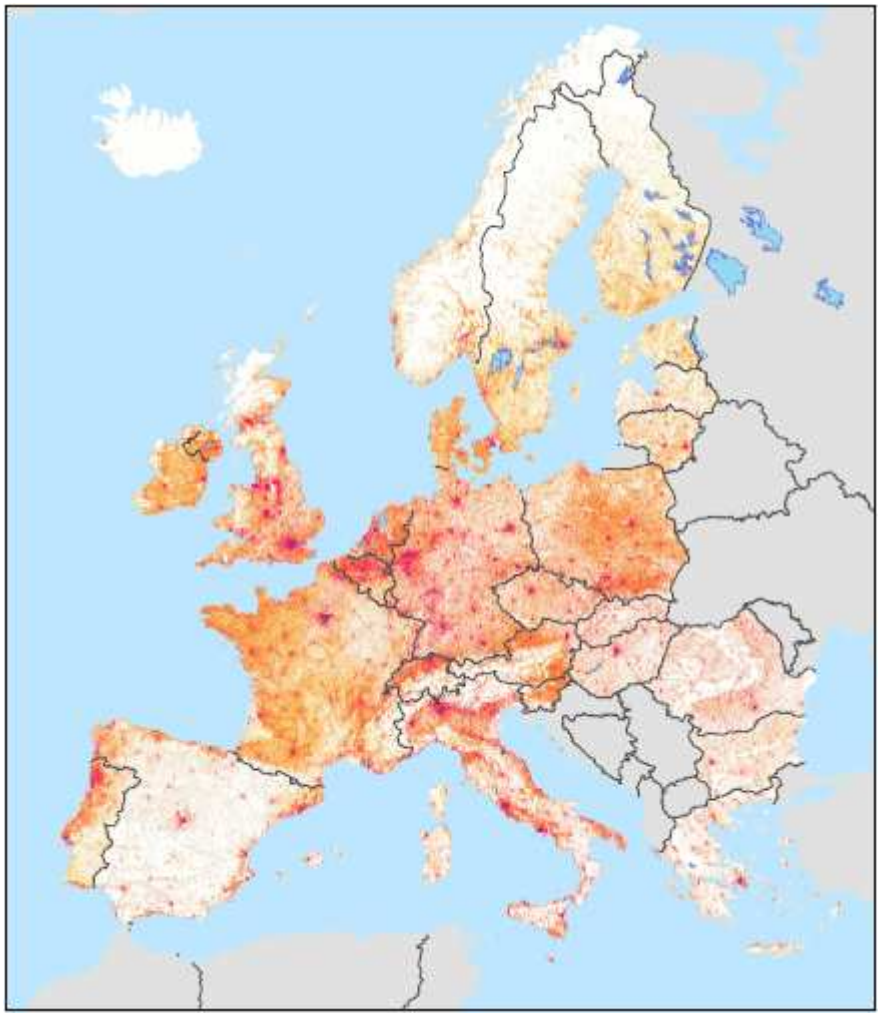
GEOSTAT 1A – Representing Census data in a European population grid



Final Report

EUROPEAN FORUM FOR GEOSTATISTICS

Version history			
V X.X	**YYYY-MM-DD**	**Creator(s)**	**Comments**
V 1.0	2011-MM-DD	Vvh, Ekp	
V 2.0	2011-MM-DD	Vvh, Niek	
V 3.0	2011-MM-DD	Vvh	
V 4.0	2011-MM-DD	Vvh, Niek, Marja, Lars, Ingrid	
V 5.0	2011-MM-DD	Vvh, Ekp	
V 6.0	2011-12-20	Vvh, Niek Ekp, Marja	
V 7.0	2012-01-18	Vvh, ekp, Marja	
V 8.0	2012-03-12	Ekp	Proofreading

Project	ESSnet project GEOSTAT 1A — Representing Census data in a European population grid	
Agreement number	50502.2009.004-2009.860	
WP	WP0 Management	
WP leader	Vilni Verner Holst Bloch (President of EFGS, Statistics Norway)	
Task		
Deliverable	1. Final Report for A 2010-2011	
Date	2011-12-xx	
Contributors	MD Mapping	Lars H. Backer
	Statistics Netherlands	Niek F.M. van Leeuwen
	Statistics France	Jean-Luc Lipatz
	Statistics Finland	Marja Tammilehto-Luode and Rina Tammisto
	Statistics Estonia	Diana Makarenko-Piirsalu and Kreet Marsik
	Statistics Norway	Vilni Verner Holst Bloch, Geir Inge Gundersen and Bjørn Thorsdalen
	Statistics Poland	Radoslaw Jablonski
	Statistics Portugal	Ana Maria Santos
	Statistics Slovenia	Igor Kuzma
	Statistics Austria	Ingrid Kaminger

Table of Contents

List of figures

List of tables

Acknowledgement

The GEOSTAT 1A project has been run by a consortium consisting of 1 coordinator (Statistics Norway), 8 co-partners (NSIs from Austria, Estonia, Finland, France, the Netherlands, Poland, Portugal and Slovenia) and 1 sub-contractor (MD Mapping).

The consortium wishes to thank the participants at the annual GISCO working groups and the EFGS conference, who have all contributed to our discussions by sharing their knowledge. We also wish to thank all those who have provided voluntary contributions, both supportive national statistical institutes and individuals. The GEOSTAT project stems from the European Forum for Geostatistics, which has a long tradition of cooperation on a voluntary basis.

A special thank you goes to Lars Henrik Backer, who has been a prime mover from the very beginning, for outlining the vision that has set this project in perspective. Thank you also to Diana Makarenko-Piirsalu and Erik Sommer, for making such a huge personal contribution, on a voluntary basis, to the administration of the website and the work on business models.

Thank you also to the participants from non-supported countries, European countries outside the EU/EFTA area, non-European countries, the JRC and the European Economic Area (EEA). Their participation in meetings and conferences shows how important the project is considered to be, and gives it a more global dimension.

And last but not least, thank you to Ekkehard Petri, our contact point in Eurostat, for all his support and understanding throughout the GEOSTAT 1A project.

We look forward to further cooperation on the upcoming GEOSTAT 1B project.

Best regards

Vilni Verner Holst Bloch

Coordinator

ESSnet project GEOSTAT 1A

Statistics Norway — Kongsvinger
Tel.: +47 / 62 88 50 97
Fax.: +47 / 62 88 52 40
Mob.: +47 / 94 82 25 32
Kongsvinger, Norway 2012-01-18

0 Executive Summary

Embedded into a long-term strategy within Eurostat and the ESS to integrate spatial information and statistics, Eurostat launched in 2010 the ESSnet project "GEOSTAT – representing census data in a European population grid dataset". The project aims at geocoding various population characteristics into a 1km² grid dataset and for this takes advantage of the 2011 Census. The present GEOSTAT 1A project has the aim to develop a vision and the the methodological foundations for a population grid dataset.

User needs have been collected on the basis of many years' experience in the countries which provide data on grids and where the grid data are used for various issues of analysis by the statistical offices themselves, other public authorities and the commercial market. Furthermore, the preconditions and possibilities for disaggregating data in the non-grid countries have been investigated.

Based on the user needs the action has developed the data specification and suggested solutions for core concerns such as map projections, the question of scales, the coding system, methods for spatial analysis, methods for delineations and confidentiality issues.

The project has described and tested disaggregation and aggregation algorithms, INSPIRE specifications and data protection rules for population grid data (both disaggregated and aggregated) and. A first set of methods to produce harmonised grid data has been put forward in this report.

At this stage the project has concentrated on total population. Ongoing development of address files and geocoding of buildings together with the census data of the year 2011 will give more opportunities to extend grid data variables.

In most countries data are published in a national grid system based on their national reference system. On a European level the Grid_ETRS89-LAEA, as defined by INSPIRE, is now accepted and GEOSTAT has therefore decided to collect data in the 1km grid of that system.

D*isaggregation methods* are used when global or European-wide population grids are made. The choice of source data is dependent on data availability and the size of the grids in the final output. Support for the development and refinement of disaggregation methods in cooperation with research institutes has been given to ensure that a full European coverage of the GEOSTAT 1A dataset can be reached.

The GEOSTAT project has examined a common process of *quality assessment.* Quality assessment has to include different approaches (geographic, statistical, production, standards, etc.) The assessment also differs according to the data sources. If grid-based data are nationally available they are produced by quite diversified methods and by different types of data sources. Documentation of the quality of data sources tends to be quite poor and production methods are just developing.

In the area of *grid-based statistics* there are two major perspectives: one is the quality of its georeferenced source data and the other is the quality of the production process. INSPIRE mainly standardises metadata descriptions and quality measures concerning georeferences and spatiality, focusing on elementary data accuracy for these topics. But the statistical world has its own standards (SDMX). The quality of grid data should be described from both perspectives.

The objective of the action has been to spread the knowledge and results gained in the project to the ESS. The EFGS website (www.efgs.info) has been used for all digital publishing. The EFGS functions as a professional reference group for the GEOSTAT 1A ESSnet project. The action has made sure that non-participating countries benefit from the experience and are encouraged to adopt and implement the approach.

During the GEOSTAT 1A project several NSIs have contributed on a voluntary basis, and many NSIs have nominated a national contact person for grid initiatives. The EFGS network now consists of national contacts from 32 countries.

To make the action sustainable, support from the ESS has been requested, among other things to solve organisational issues such as common rules for handling confidentiality and to motivate more countries to produce national grid datasets using national data sources.

Special attention is required to ensure that grid datasets can be produced not only from census data every 10 years but that the datasources are updated continuously and will allow the production of grid datasets at shorter intervals.

'One package — one provider' is the preferable way to obtain data conveniently and quickly. The project has disseminated population grid data by 1 km^2 free of charge via the EFGS web site. The goal of the EFGS is to act as a hub for users of grid statistics. *Business models* differ from one country to another. With a view to establishing 'one package — one provider', a comprehensive business model should be developed in a follow-up project. The ultimate goal remains to distribute the GEOSTAT dataset free of costs and without usage restrictions in the same way as statistics in the ESS.

No restrictions will be applied to the publication of total numbers such as the total population figure, the total number of buildings and the total number of dwellings. *Confidentiality* problems become more obvious when the amount of variables to be gridded is extended or when data are delivered by grid sizes smaller than 1 km^2. There is an urgent need for further development and harmonisation of disclosure control methods for spatial statistics by grids. Future work will include finding a solution for further variables, in particular those which break down these total numbers (e.g. population by age group and sex...).

1 Why GEOSTAT? - The Vision

'Vision without action may be seen as a daydream, but action without vision is undoubtedly a nightmare' (Japanese proverb).

Limits to growth

One of the fortunate results of the 'European project' is that our solution to the problem of production has been so successful that it is able to more than sustain the steady growth in the human population that it fuels. The challenge is that steady growth means exponential functions, and exponential functions are not sustainable.

We, the global community, have now reached a point in history where our limited reserves of non-renewable resources will sooner or later force us to apply the brake on both economic growth and population growth. We should meet the challenge by exploiting every opportunity with purposeful, well-orchestrated rational measures based on shared systems of qualified knowledge and in accordance with scientific method. This information and knowledge must be as comprehensive, detailed, frequent, and objective as possible and must receive the support of a global information infrastructure.

Responsibility in the hands of public authorities

Responsibility for defending the general interest of the global community lies with public authorities and not with a 'market' of selfish economic interests. Our governments must provide the necessary leadership and information and leave it to the market to develop the means to drive our civilisation forward. Introducing effective measures to restore our economy's natural foundation will mean setting limits to urban sprawl, restoring the natural fertility of the earth, returning to sustainable water management and implementing reforestation programmes, sustainable fisheries, etc.

Co-evolution of man and nature

Buckminster Fuller's 'Spaceship Earth' metaphor describes the earth as an integrated man-environmental system (MES). MES are too complex to be designed and built in one go. They must evolve and adapt through constant evolutionary processes which have to be carefully designed, providing a scheme for the necessary reduction of complexity. The most efficient way to reduce complexity and develop MES is the Darwin machine (DM) process that can be described as an interactive evolutionary method.

For this method to function, the following interconnected elements have to be present and work together:

- MES, a concrete system whose development is under the responsibility of a
- GOVERNOR, an institution which for its direct and indirect measures relies upon a
- GGI (Global GeoInformation system), a shared system of high-quality knowledge.

Their interaction is governed by the principle of 'learning by doing' which means that both the MES and the GGI are constantly being improved with each iteration. For this interactive approach we need to design and develop the best possible information system and develop it over time according to experience gained and changing user needs.

The GGI ('If you cannot describe it, you cannot manage it')

For it to function properly, the 'Darwin Machine' approach to the evolution of complex MES requires a reliable foundation for the GGI. The GGI contains a description of the 'narrative' consisting of operational models of the MES. These models depend on the integration of the characteristics of objects (static models) and their interaction (dynamic models).

This objective cannot be achieved without a comprehensive information model and infrastructure to feed spatial and temporal models. To characterise objects comprehensively the GGI will therefore have to support three basic information types:

- geosemantic information (e.g. alphanumeric texts);
- geostatistical information;
- Geospatial information (geographic map features).

The world of geostatistics is primarily the abstract world of 0D (point objects) geosemantic and geostatistical information, not to be confused with concrete 1D, 2D & 3D (spatial objects) geographic features modelled in the INSPIRE project. If they are to provide proper raw material for (1D-3D) spatial or temporal (4D objects) modelling and analysis, these worlds should be integrated but kept as separate components. Accordingly, all microdata should be stored in the geostatistical world of points and grids or alternative systems of regular tessellations.

Hence, geostatistical information is the foundation for all spatial descriptions of objects, just as a forest is primarily a cluster of trees, and the forest as a geographic feature is a delineation of that cluster. Similarly, raster satellite images and statistical grid datasets belong to the same geostatistical world to be processed by spatial statistics and ultimately mapped for illustration.

A vision for both the INSPIRE and the GEOSTAT projects

Although the GEOSTAT system is abstract, it is very simple. It is based on the study of points registered in space and time. Geostatistical information and point-based statistics as proposed by the GEOSTAT project constitute the informational foundation for all GGI (Global Geoinformation) systems.

The current work also aims to provide a vision for both the GEOSTAT and INSPIRE projects, primarily integrated in the building of an operative GGI that may serve as the foundation for transforming the earth into a Darwin Machine. This may serve as a vision not only for this GEOSTAT project but also for the INSPIRE project, as well as any future EU-GGI that may be developed from the contents of the INSPIRE annexes along the lines presented above.

2 Background and introduction

The GEOSTAT vision is to have a combination of grid-based statistical systems which meets needs at different geographical levels, from the national and continental level to the global level. These systems will serve different purposes, as the topics of interest vary with the geographic window in use.

The GEOSTAT action is about developing guidelines for datasets and methods to link 2011 Census statistics to a common harmonised grid. Building on the network and results produced by partners in the European Forum for GeoStatistics (EFGS, www.efgs.info), the main objective is to prepare guidelines for others to follow when producing national gridded population statistics.

Today an increasing number of national statistical institutes have the ability to produce statistics for very small areas (Figure 1). A point of departure is that NSIs usually have access to more detailed information about the spatial distribution of their population than institutes, which have made global or European efforts to estimate and model population by grids. At least during censuses most NSIs capture data by using georeferences that are even more detailed than the officially published data.

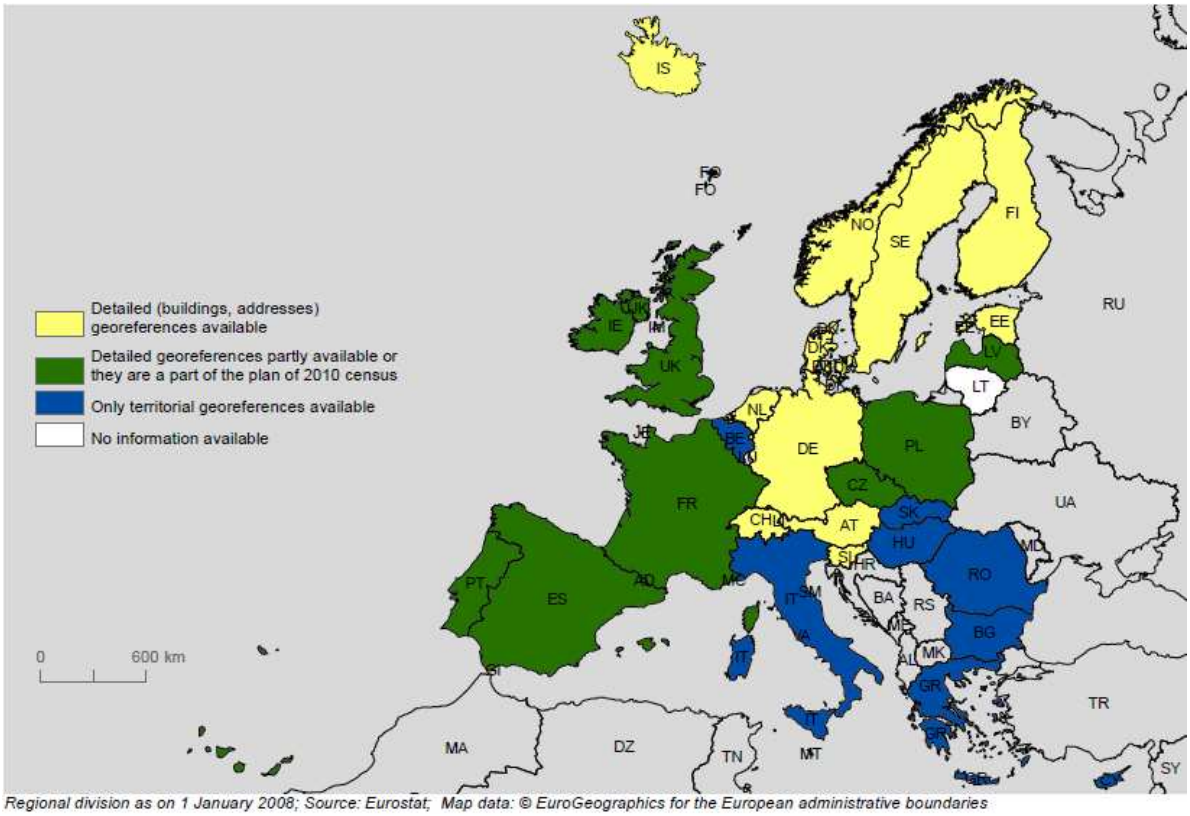


Figure 1: Available georeferences on Census 2011 data

The use of geographical grids for displaying population concentrations is not a new idea. Several early examples of this may be found in late 18th century statistical atlases. However, it is only recently that it has been possible to make population maps covering larger areas, such as Europe, in a continuous and harmonised way. In 2004 the JRC released the first version of

this kind of map, based on disaggregation techniques. However, although it does cover large parts of Europe, there are limits to the accuracy of the JRC population grid and to its possibilities for displaying further demographic variables. Hence, further work has to be done to deliver demographic maps that are more accurate and with more variables attached. The ongoing georeferencing of population census data provides opportunities for such an approach and as such the 2011 Census represents a unique opportunity to achieve very good progress here.

The main issues in harmonising various datasets from different countries are the variations in national geographic projections and grid coding systems, data interoperability between statistics and spatial data, the different time spans for updating demographic data, data quality, the range of methods for handling confidentiality and the diversity of business models and licensing policies of the Member States for statistical data.

2.1 Objectives of the GEOSTAT 1A grant

The GEOSTAT 1A project has the following objectives:

- Based on user input develop a set of methods, tools and guidelines to create harmonised data sets.
- Create - in preparation of the ultimate goal to plot the results of the 2010/11 year European censuses on km² grids - a population grid using existing population data sources.
- Prepare a vision document for a spatial data infrastructure for Geostatistics.
- Contribute to the integration of GEOSTAT with all major other major European information systems (GMES, INSPIRE, SEIS etc.) through participation in conferences, meetings, studies, prototypes.
- Disseminate and share the results from the GEOSTAT project among NSI.

3 User and producer needs and requirements

So far grid statistics have been very much a matter of national concern, and production and user needs have been based on the experiences of individual organisations or even project managers. Until now there has been no overview of the European context. In order to increase knowledge about user needs in the Member States and best practices in national organisations, two web-based surveys — Data Users Survey and Data Producers Survey — were carried out between September and November 2010. Besides learning about users' needs for grid-based statistics, the surveys included questions on the availability of grid-based statistics and related conditions of dissemination in Europe. The replies from the surveys were the starting point for formulating the requirements for a European system for grid statistics.

3.1 Survey results

3.1.1 Data User Survey

The scope of the data user survey included questions on the organisational context of the respondent, the type of studies for which grids were used, the most relevant grid cell sizes, the geographical area of interest and on acceptable usage conditions, confidentiality management and data access conditions. For more details please refer to Annex 9.1.

The answers from 45 respondents from the governmental, private and academic communities proved that the demand for grid-based statistics is high in a number of different fields ranging from the environment to telecommunications and health. This multiplicity of usage fields indicates the importance of the spatial distribution of the population in general and of grid-based population statistics in particular. According to the respondents, one of the key assets of grid-based statistics was its independence of administrative borders.

The majority of respondents used grid-based statistics for the purposes of spatial analyses. Users mainly need population, housing and economic variables on the basis of grids. The fact that grid-based statistics were available in different resolutions corresponding to the scale of the study area highlighted its significance for accurate spatial analyses. This explained the greater importance of 1km² or smaller grid cell sizes. Moreover, there was a clear correlation between area of interest and grid size. Data users who wanted the grids for areas of a country and smaller prefer the smaller grid sizes up to 250 m, while those needing the data for groups of countries or continents preferred 1km grid cell sizes.

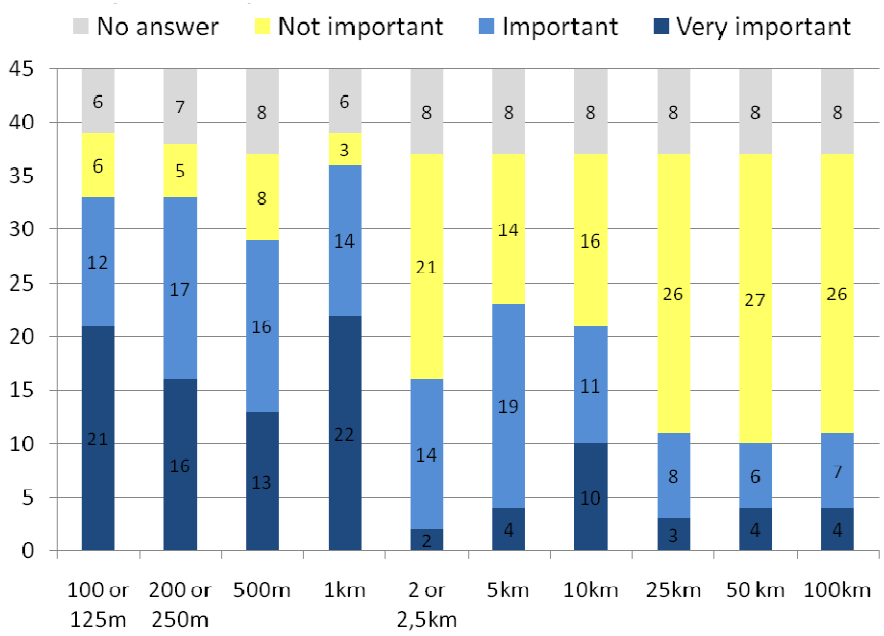


Figure 2: Importance of various grid sizes as seen by data users.

On the other hand, in the case of very small grid cell sizes users pointed to the issue of data loss due to disclosure control. In the replies there was a clear preference for disclosing absolute figures of population and buildings. In terms of the best disclosure control method, the replacement of undisclosed values with fixed minimum values, for instance, was considered more appropriate than aggregating grid cells to larger cells. As regards data access and usage conditions, users were not satisfied with the way data were made available. Therefore 'one package — one provider' was the preferred way to obtain data conveniently and rapidly.

3.1.2 Data producer survey

The data producer survey was addressed to the key stakeholders in Europe involved in the production of grid statistics. In order to match the user survey, the topics covered by the producer survey included a description of the geographical and statistical base, such as registries, the georeference plans for the upcoming census and for statistical data, grid cell sizes, projection systems, how disclosure control was managed, and how data were made available to users. The survey also asked about future plans concerning grid statistics, in particular around the census.

In all, 19 organisations replied to the survey. Of the 19 a total of 16 have produced, or are going to produce, grid-based statistics from the results of the 2010/2011 census.

According to a consultation carried out by Eurostat in November 2009 (Figure 1) and the data provider survey, the Member States can be classified in three groups according to available georeferences to be used for grid-based population statistics.

- The first group includes countries whose census data can be georeferenced by using detailed coordinates of address points, buildings or real estates.

- The second group includes countries who are in the process of adapting detailed georeferences or whose detailed georeferencing opportunities do not cover the whole country.
- The third group includes countries which collect population data by enumeration areas of different size. For these countries the census data can be georeferenced only at territorial level of variable sizes.

Most countries publish the data in national grid systems based on their national reference systems. The advantage of this is that integration with other national geolayers is ensured and grid cells are squared. However, the majority of the respondents also consider the one-grid system and Europe-wide harmonisation to be important and most agree that there should be a grid dataset available covering the whole of Europe. Here it is necessary to take into account different situations regarding the availability of georeferenced data of higher resolution, national legislation or internal NSIs regulations on statistical confidentiality. Sometimes it is necessary to maintain several grid data dissemination systems. However, most organisations would like to avoid having to maintain several systems, as their upkeep it involves additional work.

Producers rated grid sizes up to 1km as the most relevant (Figure 3) for meeting users' requirements (see Figure 2). For grids larger than 1 km the importance declines as grid cell sizes increase. Grid sizes of up to 10km are still considered fairly important, while grids larger than 10km are mostly seen as not important.

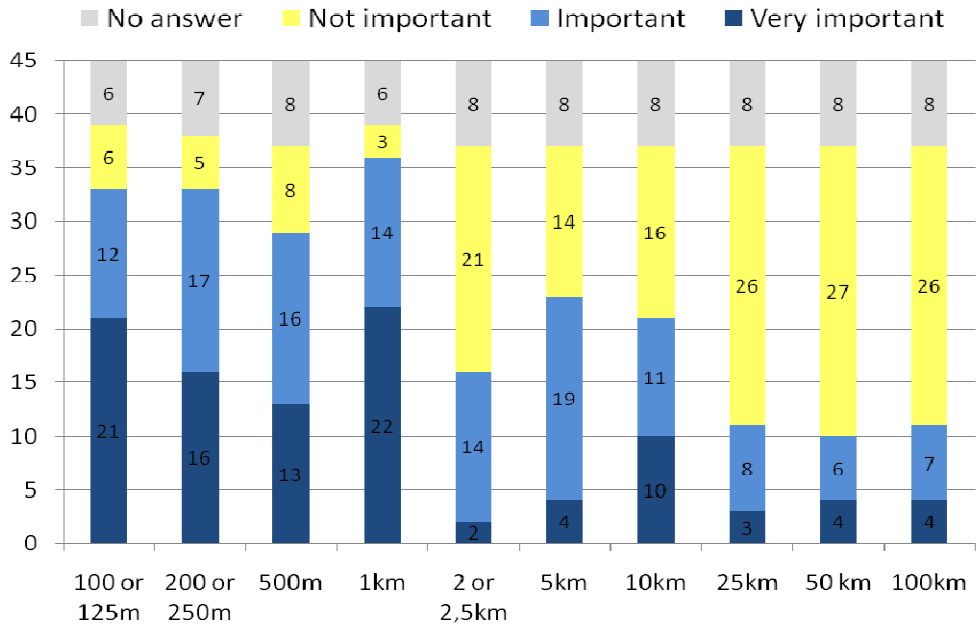


Figure 3: Availability of datasets in various grid sizes.

For reasons of data confidentiality and national data protection acts, almost all data producers face certain restrictions in their dissemination of grid statistics. Generally, it can be concluded that data providers differentiate between more and less sensitive data and define accordingly the thresholds for variables and corresponding grid cell sizes. This corresponds to availability in most grid countries, where data such as total population figures, number of buildings and dwellings are generally available on the basis of grids, for grid sizes of 1km or even smaller.

However, as a general rule, the data are released only under certain conditions, such as entering into permission agreements, making payments or referencing to NSIs. The survey also revealed various approaches to determining the smallest grid cell size for presenting a particular data variable. Various methods are in use, all of them with disadvantages — either loss of data or loss of spatial accuracy. The most frequent method for dealing with disclosure control is suppression (38 %), where undisclosed values are replaced by another number or character. Although this method results in loss of data, the grid size remains the same throughout the dataset. A proportion of 28% use the method of aggregating grids, where grid cells with confidential data are merged with other grid cells until the confidentiality threshold is reached. The resulting data layer contains grids of different sizes filling up the whole territory. The respondents acknowledged that the diversity of approaches was an issue and that European standards were required for handling confidentiality for grid data. Most grid data providers update their datasets annually.

Business models differ from one country to another. Some data are freely distributed; in other cases a fee is requested for delivering data in order to pay for the production of the statistics. After the census, grid data will primarily be disseminated in the form of tabular data formats and GIS formats as file-based data downloads. A minority still intend to publish view services. Conditions for using view services involve lower thresholds than download services. When data is to be downloaded, further restrictions, such as payment or granting permission to use by the provider, are then applied for viewing services. For 12 of the 19 data providers the distribution of data by viewing services is free of any restriction (public domain) or requires a reference to the NSI only.

3.2 Data specifications

It is very clear from both the data user and the data producer surveys that the most important requirement from public authorities and markets is harmonisation of data in terms of format, scale and access, and maximum comparability of production methods and data quality. The main issues in harmonising different datasets from various countries are: differences in national geographic projections, grid coding systems, data interoperability between statistics and spatial data, different time spans for updating demographic data, data quality measures, confidentiality issues and data dissemination policy. On the basis of the survey results, a common denominator of what is both desirable and feasible has been derived and the following recommendations can be made with regard to grid-based statistics and their publication.

3.2.1 Map projections

Projection systems in use in the Member States are optimised for the territory of the country. To match other national geolayers, national grid systems used by national statistical institutes are constructed in the national projection system. The fact that grid cells are square shaped in most cases is an advantage. As the surveys showed, harmonising map projections at the European level for all countries leads to contradicting requirements in terms of data storage and data analysis.

- Data producers do not want to maintain two data bases in different projection systems.
- Data users at the European level want a single grid system for all countries.

- Grid cells in all countries should be squared for analysis purposes.

National data producers will continue for some time to make their data available in their national systems. Generating an isolated dataset on a grid in a European system which differs from the national system will hinder integration with other data.

The adoption of a harmonised European system in addition to national systems is not possible, as simply reprojecting national grids into a European projection system will distort the formerly square grid cells. Publishing statistical data on the basis of two different grid systems, one for national users and one for European users, is often not possible as the work involved in maintaining two storage systems would be enormous and the combination of the same information in two different projection systems might lead to confidentiality leaks.

Given that national grid systems are only now being established in many countries and that INSPIRE is not fully implemented in Europe, the GEOSTAT project proposes to maintain a three-tier structure for the national, European and global levels for a transitional period. In the long term, once INSPIRE is fully implemented and automatic transformation services for all European spatial datasets are established, it might be possible to switch to an entirely European system.

The smallest grid sizes (up to 1km) are mostly applied in a local to national context. Preserving national coordinate and grid systems is acceptable here and meets the requirements of national users for squared grids and integration with other national datasets.

At the European level, the Grid_ETRS89-LAEA (Lamberth-Azimuthal-Equal Area projection) defined by INSPIRE has gained acceptance. If no direct aggregation from national point data sources into the INSPIRE grid is possible, the GEOSTAT project proposes to convert national grids into the INSPIRE 1km² grid for the European GEOSTAT dataset. A method which takes confidentiality issues into consideration is presented in Annex 9.4.

On a global scale, the European ETRS89-LAEA grid is not suitable. In Backer et al. (in print) a global system based on the UTM projection is proposed with harmonised projections and a unique coding system.

Recommendation 1: Grid data for the European GEOSTAT dataset should be referenced to the European Grid Grid_ETRS89-LAEA_1K.

3.2.2 Grid sizes

Regarding the recommended size of grid cells, 1km² appears to be a good compromise between data availability, data confidentiality and suitability for national to European study areas. The project also recommends introducing intermediate grid sizes based on a two-level quadtree (i.e. 250 m and 500 m as subdivision of the 1km grid).

Recommendation 2: The GEOSTAT grid dataset should have a grid cell size of 1km².

3.2.3 Coding system

The coding system is very important for unambiguous referencing and identification of a grid cell as well as for ensuring a clear data linkage between data and its geographic reference. The code must include all the positions of a grid and the size. A coding system requires a set of unique identifiers within each reporting area (e.g. European, national or subnational area).

As mentioned above, a three-tier system of grids is proposed. This requires a coding system for each grid system. National and subnational grid systems may continue to use their national coding systems. However, for the harmonised European GEOSTAT population grid the INSPIRE grid coding system is mandated as defined in the INSPIRE Data specifications.[1]

Recommendation 3: For unambiguous referencing and identification of a grid cell, the cell code should be composed of the size of the cell (1 and the coordinates of the lower left cell corner in ETRS89-LAEA should be used). The cell size should be denoted in kilometres ('km') for cell size 1km. Values for northing and easting should be divided by 1000. The cell code '1kmN2599E4695' identifies the 1km grid cell with coordinates of the lower left corner: Y=2599000 m, X=4695000 m.

3.2.4 Confidentiality

Most national statistical institutions are subject to their national legislation and must keep personal data confidential. This usually means that it should not be possible to identify an individual or single unit (building, employee, dwelling, business etc.). The smaller the reporting area and the lower the number of units in the area, the more important data protection is. In rural areas in particular, total counts of population or buildings are often very low in small grid cells. Disclosure control regulations to guarantee confidentiality are therefore an inevitable issue. When population data are presented on a 1 km² grid, for instance, many grid cells will only have 1 or 2 observations in rural areas.

There are several different solutions to this problem which are used in national grid systems (e.g. **Austria,**[2] **Denmark,**[3] **Estonia,**[4] **Finland,**[5] **Ireland**[6] and **Norway.**[7] See also data protection chapter), all of which are valid with different side effects.

Using methods which do not display all inhabited grid cells is considered to be rather unsatisfactory for users.

According to the user survey, methods that change the grid size and/or shapes are also thought to be impractical and less user friendly.

Basing thresholds on the number of residents can be problematic in relation to variables other than residents (households, buildings, employees). Seen in isolation it would not run counter to the confidentiality rules to say that there is only one resident within a square kilometre. However, if at a later stage it would be of interest to add other socio-economic variables, it would be more difficult, when the value is so low, to release this information when the exact number of persons in a cell already exists.

Some countries (e.g. **Slovenia**) publish absolute numbers (population, buildings, dwellings...) without restriction at any scale (e.g. 100 m grids) and the confidentiality rules are only applied to further socio-economic variables.

---

[1] See: D2.8.I.2 INSPIRE Specification on Geographical Grid Systems — Guidelines.
[2] www.statistik.at/web_en/statistics/regional/regional_breakdown/statistical_grids_etrs_laea.
[3] http://www.dst.dk/kvadratnet.
[4] https://www.riigiteataja.ee/akt/13332259.
[5] http://tilastokeskus.fi/meta/tietosuoja/index_en.html.
[6] 1993 Irish Statistics Act; No formulated disclosure rules yet, but they will involve combining a grid with an adjacent grid depending on the population and analysis.
[7] Presentation held at EFGS 2010. www.efgs.info.

For the immediate purposes of the GEOSTAT 1A project it was agreed among the project partners that the 1km² resolution can be considered to be sufficiently aggregated to be in line with all national data protection acts and that providing only the total number of inhabitants does not amount to the release of confidential information. No restrictions will therefore be applied to publishing total numbers, such as the total population figure, total number of buildings and total number of dwellings.

Future work will involve finding a harmonised solution for a 1km² grid dataset of the European 2011 census with socioeconomic variables other than just total counts (e.g. population breakdowns by age, group and sex).

Recommendation 4: For the European GEOSTAT dataset at 1km² the total population should be disclosed without restrictions for any grid cell. The GEOSTAT project recommends that data protection measures should depend on the sensitive nature of the variables. Absolute counts of the statistical units (such as number of inhabitants, households, buildings, workplaces) should be disseminated without any restrictions, even for the smallest grid sizes.

3.3 Data dissemination

In the user survey data users of cross-boundary and European datasets clearly expressed their preference for a 'one provider one package' approach. The European Forum for Geostatistics has now gained enough recognition and visibility among users to be considered the reference provider of the GEOSTAT dataset. In addition, Eurostat will promote the GEOSTAT data set on its website and guarantee access.

In line with Eurostat's free dissemination policy for statistics, all the partners in the GEOSTAT project agreed to make the 2006 GEOSTAT dataset for the total residential population at 1km² available free of charge without any restrictions on access.

European datasets are large and often heavy to process. Users may want to process the data in a software package other than a GIS. The easiest way to make the data available is in a text-based file which contains a unique spatial reference code, the INSPIRE grid of the grid cell. A spatial reference grid net will be available for creating the spatial join between statistics and geography.

Documentation relating to the dataset will be in the INSPIRE metadata schema in a separate metadata file shipped with the dataset. Where specific national usage and access conditions exist, they will have to be documented in the ‘Conditions for access and use’ field of the INSPIRE metadata standard.

Recommendation 5: The GEOSTAT dataset may be downloaded free of charge and without access restrictions in a package of a .csv file with the statistical data, a grid net shape file and an INSPIRE metadata file in ISO19139 encoding. The dataset may be downloaded from the EFGS or Eurostat websites.

3.4 Data update frequency

Most countries update their official population figures annually. A shorter time is not necessary for the type of analysis that should be borne by the GEOSTAT grid. Moreover, since disaggregation of population data will continue to play an important role in many

countries for a number of years, the reference years for the population grid will have to follow the timeframe imposed by the availability of the reference data of the most important auxiliary datasets. This refers mainly to airborne and satellite data on land use and land cover such as CORINE or the Soil Imperviousness layer, which are normally produced only every 3 to 5 years.

The reference date for the population should be the same for all countries and should be aligned with the reference date of the geometries used for the disaggregation, i.e. the census tracts or the communal boundaries. In most cases this will be the first of January in any given year. However for the 2011 Census, reference dates vary across Europe from 1 January (NL) to 31 December 2011 (SE). The error introduced by different reference dates from within the same year can probably be accepted, given the scale of the grid and the phenomena studied, which usually have a longer duration such as urban sprawl.

Recommendation 6: The GEOSTAT dataset version 1 should have the reference year 2006. The next version of the European GEOSTAT dataset should have the reference year 2011.

3.5 Data quality measures

Responsibility for data quality lies mainly with data producers. The quality mainly depends on the quality of the georeference of the register data and the quality of the statistical data, the latter being supervised and managed by the National Statistical Offices.

The quality will be documented in INSPIRE metadata, using mainly the abstract and lineage fields:

The abstract field will contain the following information:

- The very first lines should say what type of data is described and give an overall description of the data.

Further information should include, where relevant:

- General description
- Main attributes
- Legal references
- Data sources of the statistical data
- Linguistic transcriptions of the extent or location in addition to the bounding box.

Additional quality information, mainly on the data sources and the data production process, will be documented in the lineage field. It will include the following information:

- Data providers
- Production method (simple aggregation from point sources, recast from national grid system, hybrid georeferenced point data sources, aggregation/disaggregation from census tracts) and disaggregation using land cover/land use information.
- Description of the data source (registers, statistical data) with clear version identifications.

- Reference date of the population data, reference date of the spatial data (moment in time or timespan).
- Data source of the spatial data used for disaggregation (if applicable) with clear version identifications.
- Spatial reference and projection system for all spatial data.
- Update frequency of all source data used.
- Number of persons not correctly represented.

If hybrid methods have been applied, the production process has to be documented per grid cell.

Recommendation 7: The data quality of the GEOSTAT dataset version 1 should be in the form of INSPIRE metadata encoded in ISO19139 .xml files. The production process should be documented per grid data source and country.

4 The GEOSTAT 1A dataset

The main goals of the GEOSTAT 1A action are to develop guidelines for producing grid statistics and to promote national grid initiatives in National Statistical Institutes (NSI) within the framework of the European Forum for Geostatistics. However, to support the promotion task it is important to have an initial example of a dataset which already highlights the main concepts of the final data deliverable and also helps to identify obstacles and challenges. Furthermore, the dataset should be made available to users for studies and testing.

The starting point for this sample dataset was the EFGS map (ESS Eurogrid Population Map 2009) [8] of the European population on 1 km² grid. The current map is an illustration of population grid data where nationally produced population grids are integrated into an estimated European population grid (JRC 2001). The national datasets vary in terms of production process, reference date and data policy, and hence could not be freely distributed.

The consortium decided to fix 2006 as the reference year for this first GEOSTAT dataset for the following reasons:

- In 2006 European land cover and other high resolution datasets suitable for disaggregation were available.
- The year 2006 is halfway between the last 2001 census and the current 2011 census, thus providing a comparable update period.
- The 2006 data, which are now only interesting for research and analysis purposes, have a lower commercial value. For countries which sell statistical data, this makes it easier for them to provide the data free of charge.

The task was divided into two phases. In the first phase the work package partners produced population grid data for an appropriate reference year using all available data sources and testing different approaches. During the project all the partners made at least one population grid dataset covering their whole country. The national experiences gained during the production phase were collected and are now the basis for the guidelines laid down in this report. In the second phase NSIs were invited to use the guidelines and produce their own country's harmonised population grid data of 2006 to be disseminated via the European Forum for Geostatistics (EFGS) web site. The goal was to collect experiences when applying the harmonised guidelines to national datasets. At the end of the process a free and high quality dataset should be available for users.

In parallel with the NSI's efforts, the European-wide estimated population grid data were updated and improved with more accurate ancillary data and with a set of more accurate reference data by different NSIs (Steinnocher et. al. 2010, Kaminger 2011).

[8] ESS_Eurogrid Population map and Description of the ESS Eurogrid population map 2009: http://www.efgs.info/presentations/ESS_Eurogrid_Population_map_2009.pdf .

4.1 National data sources

An understanding of the nature and quality of the national georeferenced data sources that are the basis for any grid statistics is crucial for assessing the quality of the data and the limitations of national grid datasets. However, grid-based data from national sources are currently produced by quite diverse methods and from a wide variety of data sources. Documentation of the quality of data sources tends to be quite poor and production methods are just developing.

In the GEOSTAT 1A project six countries shared their experiences of producing grid-based statistics (Country reports 2010-2011). **Austria**, **Finland** and **Norway** have been producing grid data for some years now. **Poland** and **Portugal** are both developing their georeferencing during the 2010/2011 census. Until now they have had polygon-based georeferences and hardly any or no experience at all with population grids. **France** is a special case with different kinds of opportunities to georeference data linked to specific parts of the country. France's previous experience of grid-based statistics is mainly in the context of urban studies.

In **Austria and Finland** the national register of buildings and dwellings plays an important role in providing the necessary georeferences (Kaminger 2010 and Tammilehto-Luode 2010). In **Norway** this is the responsibility of the official register for ground properties, addresses and buildings (Bloch et al. 2010). Those registers provide full coverage of all the buildings in the country together with their addresses and map coordinates. A connection between a personal id-number (also id-number for enterprises, id-number for buildings etc.) to a numerical address with coordinates or centroids of buildings with coordinates allows population (and a range of socio-economic) variables to be referenced in these countries on a spatially accurate basis. Finland has produced census type data with the help of different registers since 1987. Censuses in Austria and Norway will be totally register-based first the first time in 2011.

Poland has set up a system known as **TERYT** which will provide unambiguous identification of objects with different levels of territorial detail: voivodship, county, municipality, town, village, statistical district, census region, street, building and apartment. The system contains geographical boundaries for municipalities, districts and provinces. In addition, the user address ID, which determines individual buildings in the TERYT registry, will be gradually enhanced with the x, y coordinates of the building. The introduction of geocoded address points will be a considerable improvement on the existing system of spatial identification and will enable a transition from area allocation (census areas) to point allocation. For the moment this building-level system covers only part of the country.

Since 1991 **Portugal** has had a ‘Geographic Information Referencing Base’ **(BGRI)**, which provides polygon-based georeferences for census data as far as ‘Statistical Sections’ and even ‘Statistical Subsections’. The BGRI contains only polygon features. For the 2001 Census there was a first attempt at georeferencing the census buildings within the NUTS II of Alentejo, for which the enumerators had collected the position and code number of the buildings. For the 2011 Census Statistics, Portugal has collected the coordinates of each census building. Population data georeferenced by building will be available by the end of the year 2012. However, between census years only population data by LAU1 level is available.

In France the principles of census data collection are different for municipalities with more than 10 000 inhabitants and for those with less than 10 000 inhabitants.[9] Census data is collected from small communes in the traditional comprehensive way, but geometries of census districts are not available. For larger communes census data are collected with the help of a sample of the buildings. The sample, which amounts to about 40 % of the building population, is extracted from a comprehensive building register which includes the locations of buildings and which is maintained by INSEE.

In France census data collection is a process that is spread over five years. For small communes data collection takes place once every five years but the data is not collected in the same year for every commune. For large communes the survey is done progressively during each of the five years, sampling about 8 % of the total buildings. In both processes there are no spatial considerations behind the decision regarding which entity will be sampled each year. The only consideration is that the result of the five data collections must be representative of the standard output areas of data dissemination, i.e. municipalities and sub-municipality output areas (about 2000 inhabitants). Thus, producing results on a large geographical scale from the five-year data collection and introducing spatial locations into the sample produce two different additional challenges:

- forecasting and backcasting data for five years into data for a single reference year.
- extrapolating the sample for the entire country.

Assigning locations to buildings is a particular issue for small municipalities which represent the larger part of the territory, as the data collection does not provide any information on building locations. To solve this problem INSEE, inspired by the GEOSTAT project, signed an agreement at the beginning of 2010 to obtain access to the whole cadastral register. This register was to be used as a spatial reference for administrative data (fiscal data) but only for 2/3 of the municipalities, for technical reasons. At the beginning of 2011, the availability of additional products from the French National Mapping Agency provided an opportunity to obtain full coverage of continental France.

In the meantime, non deterministic techniques were developed to import the location that was obtained from fiscal data into the individual census records, and in this way the census was georeferenced.

4.2 National production methods

In short, grid statistics can be computed using aggregation or disaggregation methods.

Aggregation is based on accurate locations of points where the data points can be added up inside each grid cell. This is the most accurate approach and provides the best quality.

4.2.1 Conversion to ETRS89_LAEA

There are specific concerns about data conversions. Point data from converted building points with the corresponding high spatial accuracy give spatially the best quality harmonised data. However, very often nationally produced data need to be converted to a different map

[9] Additional complexity comes from the oversea regions (DOMs and COMs) because of the specificities of housing there. The Geostat project is limited to continental France, so the case of the overseas territories is not discussed.

projection for the data to be harmonised with data from other countries. The recommended European grid is the ETRS89-LAEA system, which is different from most national grid nets. The tests made during the project confirm the importance of the manner in which the conversion is made. The smaller the scale of the source grid data, the higher the standard of the derived grid data can be. A study was made in Finland by converting data from the national ETRS89-TM35FIN into ETRS89-LAEA (Koivula et. al. 2011). The reference data were made by converting the source data (all inhabited buildings), before joining the points to the LAEA grid net. This method keeps the original data as they are and there are no quality errors. However, the weakness of this method is that it requires the use of primary microdata. Hence, the method produces double datasets and the national production process will have an equivalent duplicate process for a harmonised European dataset with more or less the same phases. This would make data management extremely cumbersome. Moreover, the primary datasets are not necessarily available to the grid data producer either.

Data recasting from ready-made grid datasets may offer a lighter way to reach the target. National legislation, practices and the nature of the source data determine the method that can be used and the quality that can be achieved.

Tests were made by converting 'ready-made' datasets of different grid sizes but by the national coordination system. Grid cells were first transformed into points (middle points of grid cells), then converted into ETRS89-LAEA and finally overlayed on the LAEA grid net. The differences between the reference data and the recast datasets were calculated. Comparing the 125 m grids to the reference data, the proportion of perfectly matched grids with no difference at all is 65.7 per cent of all the matched grids. The share is 52.2 per cent for 250 m grids and only 24.7 per cent for 1 km grids (see **Table 9** in Annex 9.4). This means that the grid cell size used as a starting point for the recast should be as small as possible.

4.2.2 Country reports

For population statistics this corresponds to aggregating population information from point sources such as building registers. Data can be directly georeferenced on the basis of accurate coordinates or, more frequently, can be linked to coordinates of buildings or address points using a unique ID. **Austria**, **Finland, Slovenia, the Netherlands** and **Norway** have taken this approach for the GEOSTAT 2006 dataset.

If population figures are available only for areas such as census enumeration areas, they usually have to be disaggregated using ancillary information. However, when census areas are very small and their boundaries coincide mainly in the target grid cells, fairly good estimates can be aggregated using the location of the centre point of an enumeration area. In this case the size of the grid cell should correspond to the maximum size of the enumeration areas in question (Tandem_GIS I 2002). An intermediate step towards disaggregation is a method which uses the geographically weighted population of small statistical areas. **Portugal** made a population grid map of census 2001 results based on the following assumptions:

- Whenever a polygon in the BGRI 2001 **(**a statistical subsection) is entirely within one grid cell, then its value in population is attributed to the same cell.
- When a BGRI 2001 polygon is divided by one or more cells in the grid, then the value of the resident population of the polygon BGRI 2001 is distributed according to the percentage of area that is in each cell.

The availability of the geo-referenced 2011 census buildings allowed Portugal to develop further their production method on the basis of this information. At this point only population data for each statistical subsection is known, so that the method is based on the assumption that each residential building has the same number of individuals.

The population for each grid cell (G) is given by

$$G_j = \sum_{i=1}^{N_j} S_{ij}$$

where:

$$S_{ij} = Population\ of\ subsection\ i \times \frac{Number\ of\ 2011\ residential\ buildings\ within\ 2011 subsection\ i\ in\ grid\ j}{Number\ of\ 2011\ residential\ buildings\ in subsection\ i}$$

$$N_j = Number\ of\ subsections\ in\ grid\ j$$

The exact population for each building will be known by the end of 2012. It it will therefore be possible to determine the exact population within each grid-cell for the 2011 census (Statistics Portugal 2011). Population data for the non-census year 2006 is only available at municipality level (LAU2). The 2006 population for each grid was estimated using the 2001 and 2011 population distribution based on estimated population data for each municipality (LAU2).

In small municipalities **in France**, where data collection is comprehensive, simple aggregation of individuals provides the population counts, but only for the reference year of the data collection for each commune. Unfortunately the reference year differs from one commune to another. As a result the published total municipal population must be used to forecast or backcast the aggregated data for a specific reference year and to respect the total national population for that reference year.

For large municipalities the census results are extrapolated to each building, even if it was not sampled. The computation is done using spatial autocorrelation techniques applied to the neighbourhood of each building.

In Poland the address register does not yet cover the whole country. For testing purposes grids were produced based on population data from both disaggregation and aggregation methods. The analysis was made covering one subregion, Piotrkowski. For the disaggregation method, population data were analysed by smallest available area and soil sealing (SSL) data layer with 100 meter resolution.

The spatial disaggregation method:

$$PGK = \Sigma K\ PssL$$

$$PssL = Pop(N) / M$$

PssL = Average value of the population at the point on Soil Sealing Layer for N-th of the census district, in which there is 'M' points of SSL.

PGK = Total population assigned to points from Soil Sealing Layer located in the space grid 'K'.

To verify the results obtained from the test area the thematic map was produced using an aggregation method. The source of data about the population was the statistical address points layer formed for national census needs. Differences in the results by the different methods were studied visually and by calculation quality indicators. Using the experiences gained from the tests of one subregion, Poland produced population grid data covering the whole country by a hybrid approach. An additional challenge was that the address point layer used was from

the year 2011 and the target population data is from the year 2006 (A State of the Art report – Poland 2011). The 2006 population was estimated by applying a weighting corresponding to the municipalities' population gain/loss during the period 2006-2011 (Jablonski 2011).

In **Estonia** the population data for non census years is only available at municipality level (LAU2). In preparing for the 2011 census a number of datasets were created such as building centroids (containing building addresses and coordinates), individuals' address data from the population register, and data about new buildings from the buildings register. Combining the official 2006 population counts at municipalities level with those data allowed to accurately estimate the 2006 population at grid level.

The addresses of individuals and buildings were matched, yielding a building centroid dataset with the number of persons in each building. The building centroid dataset was combined with data from the buildings register and the buildings that were built between 2006 and 2011 were excluded from the centroid dataset. On the basis of the building centroids the 2006 population in municipalities was calculated and the numbers were compared to the official population counts for the year 2006 at LAU2 level. If differences were found between the two numbers, the missing/redundant number of people were added to or deleted from randomly selected buildings within the same municipality. Hence, the building centroid datset could be used as a microdata input dataset for aggregation to the grid cells.

4.3 European grid production initiatives

In parallel with national efforts within the GEOSTAT project to produce the best possible grid data for each country, European dasymetric maps were improved. Disaggregation of the European population is a modelling approach and therefore outside the scope of the GEOSTAT project, which aims at strengthening the statistical method of producing population grid datasets.

The starting point for an improved disaggregation approach was the experience from the JRC 2004 dataset based on population 2001. The disaggregation of NUTS3 or communal population data and the limited spatial resolution of the CORINE 2001 Land cover dataset have lead to an underestimation of built-up areas and an overestimation of sparsely populated areas, at least in European-wide dasymetric maps (Gallego 2010, JRC 2001).

The relatively recent dataset GMES Fast Track Service Precursor on Land Monitoring, which provides the degree of soil sealing for EU27 + countries, has helped improve spatial disaggregation methods for the whole of Europe.

This new disaggregated European population grid contains total population figures for the reference year 2006, disaggregated to a 1km² grid. The dataset has been produced by the Austrian Institute for Technology (AIT[10]). The data cover EU 27 and EFTA countries (except Cyprus, as there are no LAU population figures for CY). The source population data is reported on LAU2 (LAU1 for PT). Since the degree of soil sealing does not correspond directly to residential building density, a number of pre-processing steps were performed beforehand masking potentially inhabited land. For evaluating the disaggregation results, reference population grid data of several register countries were used. (Kaminger 2011, Steinnocher 2010). Further references and exhaustive documentation on dasymetric

[10] http://www.ait.ac.at/.

population mapping of population grids can be found in the literature indicated in the references in this section.

In addition to the European disaggregated grid AIT also produced national grid datasets for those countries that share a border with a GEOSTAT country. This simplifies the integration of the two datasets along the borders and to deal more accurately with population in shared border cells.

4.4 Description of the GEOSTAT 1A dataset

The data harmonisation work included a first iteration of the collection of concrete datasets from National Statistical Institutes. The first harmonised dataset contains only population counts by 1 km² grid cells, preferably for the reference year 2006 (see 3.2).

The proposal is to collect data per counrty and make it available via the EFGS web site. The data will then be integrated with the AIT 2006 dataset for the missing countries into a first version of a GEOSTAT dataset. This first version will be improved in subsequent iterations.

The starting point was a European-wide grid net of 1 km^2 cells corresponding to the INSPIRE specifications[11] and covering EU27 + EFTA countries. The grid net was made with an ArcGIS extension and can be obtained free of charge from Eurostat at ESTAT-GISCO@ec.europa.eu. Finally, the grid net was divided into country nets allowing for easier data handling.

Those grid nets represent the framework for the integration of national grid data. The actual grid dataset consists of .csv textfiles with the unique INSPIRE grid cell code as reference.

When conversions of national data to the INSPIRE grid projection system were tested, the quality of the data proved to be a source of concern (see Annex 9.4, Koivula et. al. 2011). The preferred option, ensuring the best quality, is always to aggregate directly from point data which are in the same projection system (ETRS89-LAEA) as the target grid. For those countries which were not in favour of direct aggregation from point data sources in the ETRS89-LAEA projection, the proposal was to use data via the smallest available grids and recast the data to the INSPIRE grid.

The conventions for naming the national grid dataset file and the variables were as follows:

Name of the file	GEOSTAT_grid_POP_1K_CC_YYYY (CC: country code, YYYY: ref. year)
Column Names	GRD_ID, METHD_CL, YEAR, POP_TOT

Where:

GRD_ID	Identification code of the grid cell (lower left-hand corner) according to INSPIRE
METHD_CL	Method used for the grid cell; A (aggregated), D (disaggregated) and M (mixed)
YEAR	Reference year of the data
POP_TOT	Population count of the grid cell rounded to integers (in the case of border cells the cell

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

	contains the sum of the original cells).

Cells with zero total population were removed from the dataset.

Regarding the METHD_CL attribute for each grid cell, there can be various combinations of actual methods combining aggregation, disaggregation and even other estimation methods. For the sake of simplicity the goal is to differentiate cells which have been merely disaggregated using a dasymetric approach ('D'), from those grid cells which are simply 'point in polygon counts' based on detailed georeferenced source data ('A') and from those which apply various methods and data sources to estimate and model the population figure ('M'). For further details metadata will be used.

Metadata were encoded in XML files applying the INSPIRE standard for spatial metadata and using the EU Geoportal metadata editor.[12] For the moment, the tool provides a rough description of data from a geographical point of view. There is hardly any room for descriptions from a statistical point of view, e.g. describing the structure and quality of the data in SDMX. For future iterations further investigations are required into how to integrate good quality statistical information into the metadata of the dataset and how detailed this information has to be, in terms of lineage information and production steps of the dataset for instance.

By the end of the project, in addition to the nine partner countries, Denmark, Sweden, and England and Wales volunteered to deliver their population data by harmonised LAEA89_1K grids from the year 2006 via the EFGS web site.[13] In addition, national population data by 1km² grids are available in Kosovo, Northern Ireland and Switzerland (reference year 2001) and Ireland (reference year 2011) as well as Spain (reference year 2006 and 2011) produced by various methods. National grid data from other years, or containing additional variables, or even by other grid sizes, are made available via the EFGS web site by France, the Netherlands, Norway and Portugal.

Besides making national datasets available, the aim was also to produce an integrated GEOSTAT dataset for EU27 + EFTA. Hence, for those countries which did not contribute with a national population grid, the dataset will be filled with data from the AIT 2006 dataset. This concerns Belgium, Bulgaria, Germany, Slovakia, the Czech Republic, Hungary, Latvia, Lithuania, Ireland, the United Kingdom (Scotland and Northern Ireland), Luxembourg, Liechtenstein, Switzerland, Italy, Malta, Romania, Greece and Spain.[14]

The final European GEOSTAT dataset will therefore contain two additional attributes:

CNTR_CODE	ISO code of the country in which the grid cell is located (in the case of border cells the country codes are concatenated) and the national value is provided in brackets, e.g. SE(7):NO(14).
DATA_SRC	For national datasets the country code; for the European dataset 'AIT'; in the case of border cells from different sources, the DATA_SRC

[12] http://www.inspire-geoportal.eu/index.cfm/pageid/342.

[13] http://www.efgs.info/data/geostat/open-data.

[14] For Spain a disaggregated grid is available produced by Francisco Goerlich from the University of Valencia using the national SIOSE Land Use/ Land Cover data as ancillary data. It might replace the AIT data for Spain at a later stage.

	codes are concatenated, sperated with a colon, e.g. SE:NO.

Along the boundaries where datasets from different data sources are joined the population figures from either source are added up and transferred into a single joint grid cell as illustrated in Figure 5.

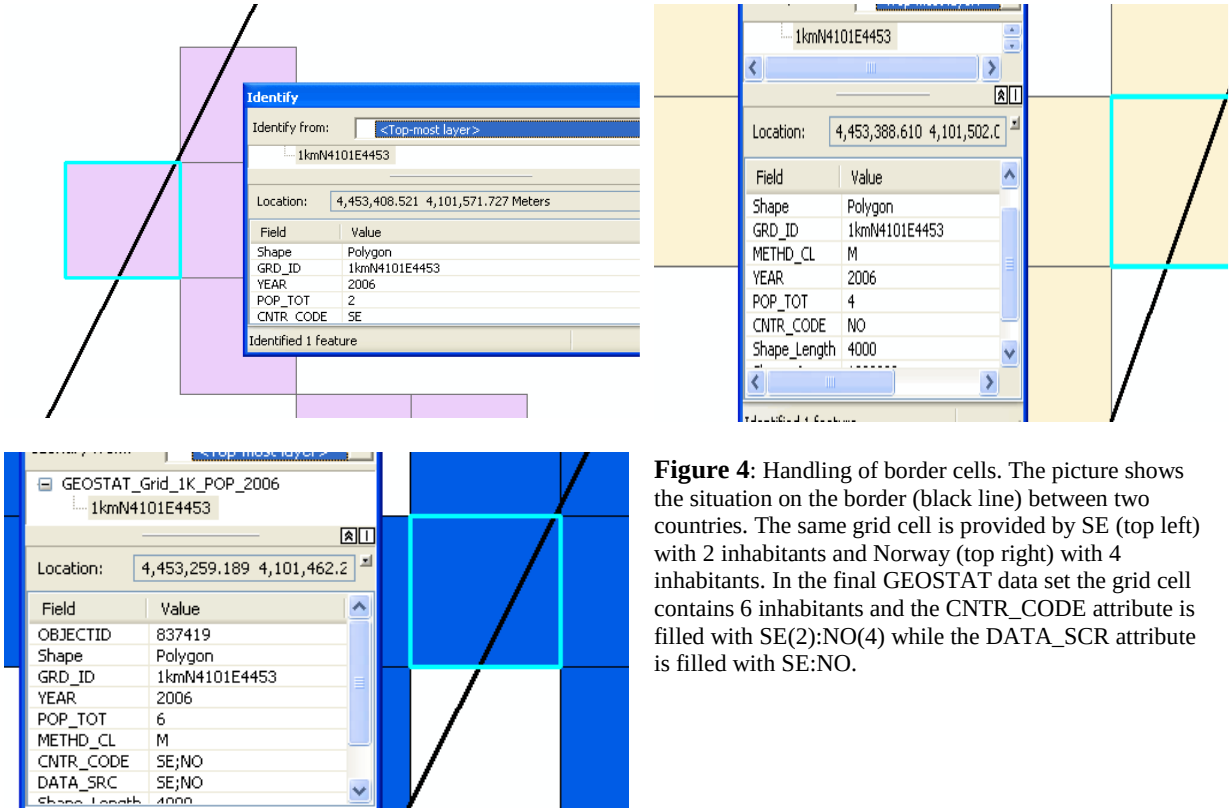


Figure 4: Handling of border cells. The picture shows the situation on the border (black line) between two countries. The same grid cell is provided by SE (top left) with 2 inhabitants and Norway (top right) with 4 inhabitants. In the final GEOSTAT data set the grid cell contains 6 inhabitants and the CNTR_CODE attribute is filled with SE(2):NO(4) while the DATA_SCR attribute is filled with SE:NO.

4.5 Statistics of the GEOSTAT 1A grid dataset

The reference gridnet Grid_ETRS89_LAEA_Europe_1K[15], intersecting with the EU27 + EFTA landmass[16] delineated by a coastline/boundary dataset at scale 1:100 000, has a total of 4 884 516 grid cells whereas the GEOSTAT_2006_POP grid contains 1 945 315 inhabited grid cells with at least one inhabitant. This means that ~40% of the grid cells intersecting the EU27 + EFTA landmass are actually inhabited.

In total the GEOSTAT national dataset represents 43 % of the total 505 Mio EU 27+ EFTA citizens population in 2006 and covers 2 446 199km² equal to ~50 % of the total land surface of 4 825 275km² while the disaggregated population dataset covers 57 % of the population.

The avarege population density in the GEOSTAT territory is 86 inhabitants per km² whilst the average number of inhabitants per inhabited grid cell is 170 per km².

While there are ~70000 grid cells with only one inhabitant of which 20 774 are drawn from population registers, the highest observed population per grid cell corresponds to 52 898 inhabitants and is located in the centre of Barcelona.

[15] http://www.efgs.info/data/eurogrid/Grid_ETRS89_LAEA_Europe_1K.zip/view .

[16] This includes also territory covered by inland waters and lakes.

4.6 Mapping the dataset

The combination of the national datasets and the AIT dataset results in a European-wide dataset that can be mapped using an INSPIRE grid net of 1km² cell size. The following population classes are recommended for general mapping purposes:

1. Not inhabited, Population 0, Colour scheme CMYK (white)
2. 1 – 4 Inhabitants, CMYK M5, Y40 (light yellow)
3. 5 – 19 Inhabitants, CMYK M20, Y60 (yellow)
4. 20 – 199 Inhabitants, CMYK M50, Y80 (light orange)
5. 200 – 499 Inhabitants, CMYK M70, Y80 (orange)
6. 500 – 5000 Inhabitants, CMYK M100, Y50 (red)
7. >5000 Inhabitants, CMYK C55, M90, Y50 (lilac)

The GEOSTAT 1A map follows the classification and colour scheme of the 2009 ESS Eurogrid Population Map. This classification of the population counts by grid cells was chosen after looking at different national practices. The choice of colours was for maximum differentiation between classes of relatively small grid cells.[17]

On the basis of the above classifications the following frequency distribution is found in the Grid dataset:

Table 1: Frequency distribution of population classes in the GEOSTAT 2006 dataset (number of grid cells per class).

0	1-4	5-19	20-199	200-499	500-5000	>5000
2 939 201	275 784	519 129	786 903	172 788	179 214	14 072

Depending on the purpose of the mapping class, different thresholds and colours may be chosen. Another possibility is to include the threshold of urban clusters used in the degree of urbanisation proposal by Eurostat (Eurostat 2011). One point of departure might be to colour each grid cell on the map, but there are plenty of different other techniques to be tried out in the future.

4.7 Quality assessment of the dataset

Given the diversity of national data sources and production methods of the 2006 dataset, quality assessment is mainly a national exercise. The topics include geography, statistics, production process and application of standards.

Under the assumpation that national figures have better quality and represent the reference to validate against, the most straightforward approach is to offset national results against a European-wide dasymetric map by:

[17] http://efgs.info/presentations/Description%20of%20the%20Population%20Grid%20Map%20of%20Europe_13022009.pdf .

- comparing for each grid cell the number of inhabitants grid cells between estimated data and register data;
- computing the Total Absolute Error (TAE) for all grid cells.
- put the TAE in relation to the total population of the area.

The error per grid cell is defined as:

$$R_s = \widehat{P}_s - P_s$$

whereby $\widehat{P}$ is the reference population of a grid cell s and P is the estimated population of the same grid cell. The TAE is defined as:

$$TAE = \sum_{s=1}^{n} \left| R_s \right|$$

And the relative Error:

$$relError = \frac{TAE}{\sum \widehat{P}}$$

with $\widehat{P}$ as the reference population of a given territory.

A quality assessment of the disaggregated population grid carried out by AIT showed that the the relative error at national level for those countries where reference register data are available is in the order of 30% but can exceed 50% for sparsely populated countries such as NO, SE and FI.

Dasymetric maps are more successful in detecting whether grid cells are not inhabited at all. Here the share of cells which are correctly classified as not inhabited is between 85% and 95%.

On the other hand for inhabited grid cells the detection factor is only between 63% and 95%, depending on the country with the lower rates in the sparsely populated countries.

Finland made a comparison with the JRC grid data (Gallego 2010) (last row of Table 3.1) testing the effects of converting national data into the harmonised ETRS89-LAEA data. The primary purpose was to compare the results of conversions of the source data (first row of Table 3.1.) with the direct conversions of ready made population data of different grid sizes (recast data - rows 2 – 4). Table 3.1 shows differences between national (conversion of source data) and recast population grids together with the results of an estimated population grid by the JRC in Finland in 2011 (Appendix 9.4: Koivula et.al. 2011).

The results show that all the recast data sets yield very good results in terms of number of inhabited grid cells (deviation max. 3% for the 1km² source and total population) in comparison with the estimated (JRC) dataset. The total population is also quite well preserved in the JRC grid, although it is distributed over a much higher number of grid cells, thus losing out significantly on spatial accuracy. However, there were also significant differences in the distribution of recast datasets, depending on the size of the grid cells in the dataset (see Koivula et al. 20111). The differences are mainly found in the lower TAE classes (1-10 persons). As a result of the intrinsic nature of the way the grid cells are recast, the range of error is limited to the surrounding grid cells: Max. Range = $\sqrt{2km^2}$.

Table 2: Quality assessment of recast population grids and disaggregated grid. Line 1 contains the aggregation of the original point data which was projected to ETRS89-LAEA and then aggregated. This is the benchmark for lines 2-5. Lines 2–4 contain statistics on grid cells originally in ETRS89-TM35FIN and then recast to grid cells in ETRS89-LAEA. Line 5 contains statistics on the JRC2004 dataset. N= number of inhabited grid cells, Mean = average inhabitant/grid cell, SUM= total number of inhabitants, Minimum= min. number of inhabitants of all grid cells, Maximum = max. number of inhabitants of all grid cells

Dataset	Number of grid cells	Mean	Sum	Minimum	Maximum
Dataset from converted building points (the reference dataset)					
1KM_ETRS89-LAEA	102 050	51.0	5 204 192	1	14 053
Datasets from converted grid points (by recasting)					
1KM_ETRS89-LAEA from:**125m** ETRS89-TM35FIN	102 249	50.9	5 204 192	1	14 197
1KM_ETRS89-LAEA from:**250m** ETRS89-TM35FIN	102 759	50.6	5 204 166	1	13 283
1KM_ETRS89-LAEA from:**1km** ETRS89-TM35FIN	99 049	52.5	5 204 179	1	19 175
JRC disaggregated dataset					
JRC_DISAGG	159 921	32,4	5 181 806	0,01	5 866

The tests were made on Finnish datasets exclusively. It would also be interesting to carry out tests and comparisons with datasets from other countries.

France also made a comparison between JRC population grids and France's own estimation of population grids (see Figure 5).

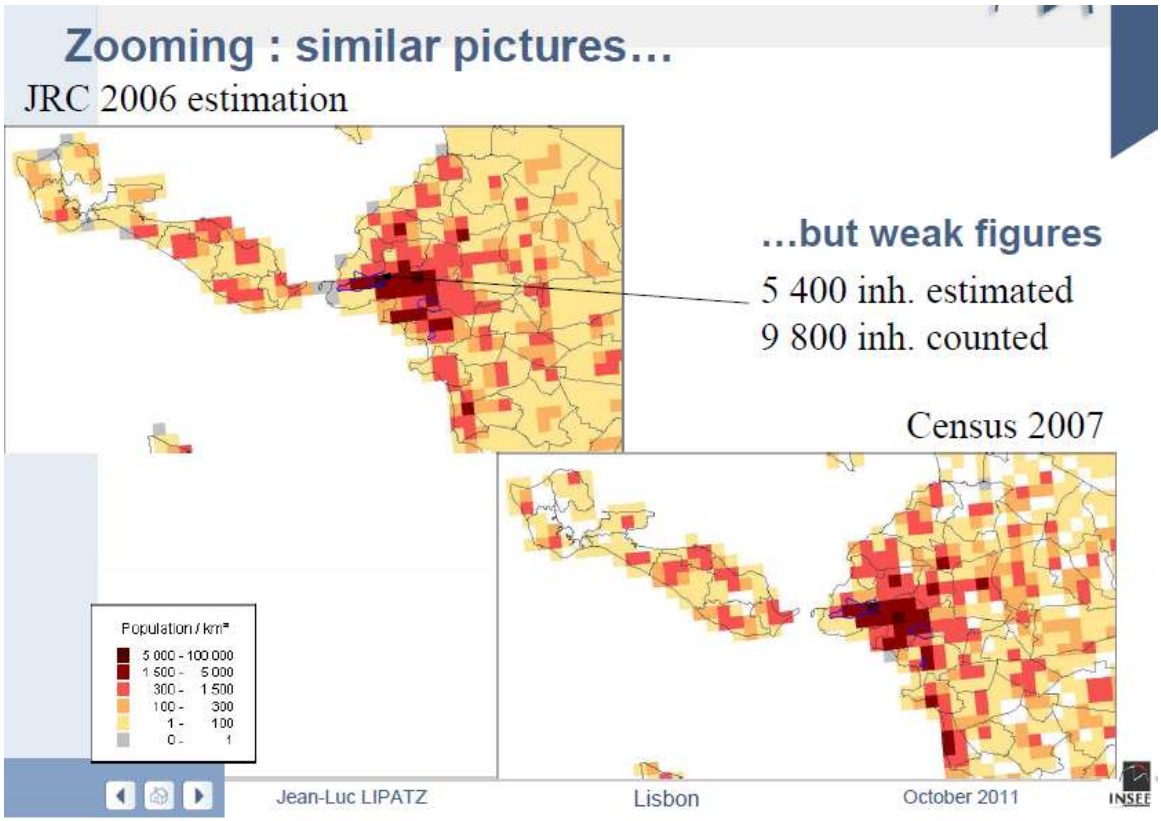


Figure 5: Comparison of JRC population grid with French national population grid.

While a purely visual inspection gives the impression that both datasets are comparable, the actual figures differ by nearly 50 % for the small sample area.

Like Finland, France also argues that changing the coordinate system or using different size grids also gives different figures.

A relevant approach for quality assessment is to compare the results of the gridded population with the official statistics already published at higher territorial level (e.g. communes).

France used a quality indicator calculated by comparing an uncorrected prediction of population ($\hat{P}$) in each municipality to the actual value (P). The figure gives an indication of confidence in the process. A weakness indicator was computed for each NUTS3 by:

$$\frac{\sum \left| P - \hat{P} \right|}{\sum P}$$

The results show that this indicator ranges from 2 % to 98 % but that most of its values lie in the range of 20 %-30 %. The problem with this kind of measurement is that it mainly affects municipalities with small populations for which an estimation error of 100 persons may represent an extremely large relative error.

Poland made comparisons in a test area between a population grid layer made by data from census districts proportionally assigned to the points of soil sealing layer (SSL) and a grid layer made by assigning population to real address points. The comparison showed that the disaggregation method gives better results in urban areas than in rural areas. It was also discovered that disaggregation results may be improved by using additional information/estimates about areas with no inhabitants; water areas, wetlands, large forests etc. by excluding these areas from the datasets.

In the register countries quality assessments are made for source data by estimating their spatial coverage but mainly by comparing data at macro and micro level (see **Table 10**). For example, population census data are compared with relevant sample survey data (e.g. data from the labour force survey and from the survey on living conditions and housing). Outliers such as population (buildings) on the water or outside the country's borders are also detected.

There is still no standard form for documenting quality. For this project phase the INSPIRE metadata standard has been used to document data lineage. More detailed information was included in the form of links to further references. Nevertheless the INSPIRE standard is not suitable for exhaustive quality documentation and another format will have to developed for future iterations of the GEOSTAT dataset (see example in section 6.2.4).

4.8 Further development issues

During this first iteration of the GEOSTAT dataset detailed experiences and insights have been gained on how best to use the source data, generate the grid dataset from the source and evaluate and document its quality. Naturally, in this initial phase many potential improvements have been already identified which will be addressed in future iterations. Topics range from more comprehensive and detailed metadata, through standardised quality measures to a regular update procedure for the data sources. One of the first major

improvements to the actual data quality will be the introduction of 2011 census data into the GEOSTAT dataset.

At this stage the project has focused only on total population data. However, there are already grid data available relating to many other variables (e.g. population by age categories or workplaces) mainly in those countries that use registers and aggregation methods (**Austria**, **Norway**, **Sweden**, **Finland**, the **Netherlands**, **Denmark** and **Slovenia)**.

France has also been producing these data for several years, but to a restricted extent geographically. There are also other recent or ongoing projects targeted at producing other new variables by grids.

The ongoing development of address files and the geocoding of buildings, combined with the census data of the year 2010/2011, will give more opportunities to extend the range of grid data variables or the scale of geographical grid cells.

The development of refined disaggregation methods along with more and better ancillary data in cooperation with research institutes and national statistical institutes will also help to improve disaggregation methods. It remains to be seen to what extent the disaggregation of further variables will yield meaningful results. Compared to register-based data sources and aggregation, disaggregation is only the second best option and will always have serious limitations in terms of further statistical variables beyond total population.

In many countries census data provide more opportunities to produce small area statistics such as grid data. Given that a census is carried out only every 10 years, improved methods for estimating the figures between censuses or making grid data from old data with poorer georeferences may also be issues to be considered and developed further in the future.

Conversions of grid data from national projection systems to European projection systems were tested with different sizes of national grids. The conversions were made by taking the centre points of grid cells as reference points. One possible refinement could be to calculate a population-weighted centre point. There might also be other ways to control the conversion process (e.g. using interpolation of source data).

The side effects of harmonising different national datasets have not yet been discussed in depth. At a technical level, joining grids together and handling overlapping grid cells from different sources will have to be supported by GIS tools.

The role of EFGS as a data provider and the definition of a common dissemination policy, at least for part of the grid data deliveries, are issues that need to be agreed in the future. The role of the INSPIRE model of decentralised data delivery via open interface must be discussed as well.

Finally, a common approach to disclosure control remains high on the agenda and will have to be addressed, at least for the basic demographic variables.

5 A method for generating population grid statistics

The present GEOSTAT 1A project represents the starting point in an iterative development, the ultimate goal being to produce a high quality population grid of the 2011census. The prototype dataset delivered is an excellent starting point for future iterations but has certain limitations such as a fairly diverse production processes at national level; and the fact that it represents only total population counts at theplace of residence. Looking ahead, the following section presents a blueprint for all aspects of the final census dataset and also, where decisions or information are still missing, points out the issues at stake that still require decisions and agreements which are often beyond the technical level of this project.

5.1 General method to generate grids

'A grid for representing thematic information is a system of regular and georeferenced cells, with a specified shape and size, and an associated property' (European reference grid 2003).

The production of statistical grids usually consists of two elements: a grid net and a statistical table. A grid net is a vector layer of regular grid cells covering the whole area in question. It can be made with almost any GIS tools. The grid net represents a steady and interoperable location of different statistical objects. Above all, it can provide a basis for the integration of different thematic data. The European harmonised 1 km² grid net is available in its entirety and by country on the EFGS web site.[18]

A grid net is not always necessary, because georeferenced data can be processed just like any territorial statistics or like spatial statistics by point georeferences. Grid-based statistics are territorial statistics which use direct x and y coordinates as a location code. For a regular grid net the x and,y coordinates usually denote the lower left corner of a grid cell of a given size.

The first precondition for generating grid-based statistics is that thematic data can be georeferenced. This means that the data units have direct or indirect links to geographic coordinates.

If there are links available to coordinates of buildings, address points or real estates, the method to be used is called the aggregation method and in GIS terms uses a point in polygon procedure to aggregate the data.

If the data is available only by territories (e.g. enumeration areas) the precondition for grid generation is that the boundaries of the areas, or at least the centroids of the polygons, are available as an input to disaggregation.

5.1.1 Aggregation method

The overall workflow of data aggregation is illustrated in the following Figure 6.

[18] http://www.efgs.info/data/eurogrid.

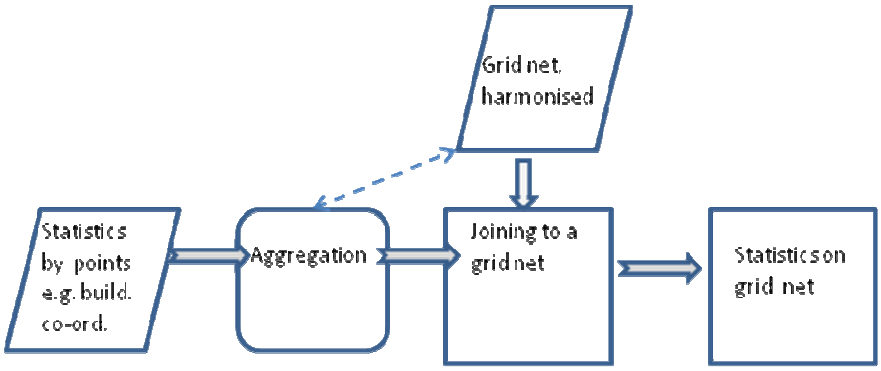


Figure 6: Workflow for spatial grid-based statistics by the aggregation method

There are several alternatives for conducting an aggregation process. The whole process may be done using statistical software or with the help of GIS software, or a combination of both. The coordinate system and projection used should be the same for the grid net and the data georeference and should remain so during the production phases. If required, a data transformation stage has to be carried out first. The conversion from one projection system to another, if needed, is best done with the source data. However, data which is already gridded can be converted to another coordinate system, although this will result in small changes to the data. In this case it is recommended to recast the data from the smallest available grid size to minimise the impact (Koivula et.al 2011).

5.1.2 Disaggregation method

The overall workflow of the disaggregation approach is described in detail in the literature (see Batista e Silva et. al 2011, Gallego 2010 for further references). Although refining a European disaggregated grid was outside the scope of this project, such a grid will be fundamental to producing a grid map for the entire European territory for a period of some years, until such time as national grid initiatives are fully established.

The disaggregation method is particularly useful when large territories are covered in a multinational to global context. The choice of source data depends on data availability and the size of the grids in the final output. Spatial statistics by census enumeration areas are usually the best starting point. If the area of statistics is greater than the target grid cell, ancillary data should be used to predict the distribution of statistics according to corresponding grids. For population grids, various ancillary data are available to describe indirectly where there is population (see Figure 7). A disaggregation model is needed to estimate, for each grid cell or part of grid cell, different density categories defined by ancillary data. The calculation of density categories and weights depends on the ancillary data available (Gallego 2010, Batista et al 2011, Goerlich et. al 2011).

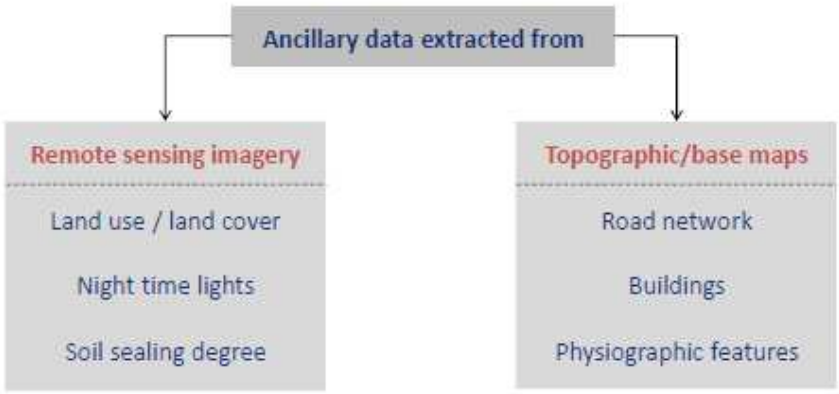


Figure 7: Examples of ancillary data for estimation of population in grid statistics.

A specific case of disaggregation is where enumeration areas are so small that they fit inside the target grids. In these conditions the aggregation method can be applied (see Corcoran 2011 for more details).

National disaggregation initiatives often have access to high-resolution ancillary datasets that are not available with comparable quality throughout Europe (Goerlich et. al 2011). Some countries have used information about the location of buildings to predict the distribution of population inside the grids (Corcoran 2011, Schoenmakers 2011). This means that when the enumeration area (with a known population) is split across grid cells, then the population of a grid cell depends on the percentage breakdown of address points within each enumeration area.

5.2 Confidentiality management

Handling confidentiality is one of the biggest challenges of grid-based statistics. When the size of a grid cell is small or a certain number of variables per grid cells are released, there is a direct and indirect risk of disclosure of statistical units. Although EU and national legislation on data protection of statistical results are similar, the actual implementation of general rules to small-area statistics differs from country to country.

The GEOSTAT project recommends that data protection measures should depend on the sensitivity of the variables. Absolute counts of the statistical units (such as number of inhabitants, households, buildings, workplaces) should be disseminated without any restriction, even for the smallest grid sizes.

However, if the grid cells are not sufficiently populated, further characteristics of the statistical units should be protected (e.g. a person's age). A data protection threshold should depend on the type of statistical unit. Characteristics relating to people should have greater protection than those relating to buildings, for instance.

One way to control the dissemination and use of grid-based statistics is to release them on condition that a specific licence is granted. The licence will allow the data to be used for restricted purposes (such as scientific research or statistical surveys on society) and by authorised institutions or persons only.

Issues relating to data disclosure have also been a subject of concern when data are produced in more than just one coordinate system or projection. However, using the recasting method

(Koivula et. al 2011) mitigates this risk. The original resolution and protection levels determine the data disclosure limitations. This means that if in the national coordinate system (and projection) of 250 m x 250 m data are safe from the point of data protection, no extra information is revealed after conversion to the grid middle points. The situation is different when the original primary dataset is actually converted (e.g. building points in the Finnish dataset). In that case a comparison of the converted and national datasets may expose highly detailed information, which may also contravene the disclosure control rules in place.

The most commonly used disclosure control method is suppressing the confidential cells. However, this has the effect of making the data less useful for analysis. Rounding (Masik 2011), scrambling (Lipatz 2011), clustering (Sommer 2007) or using different grid cell sizes (Sehlin 2011) can also be used, depending on the use to which the data will be put.

An agreement on the use of similar confidentiality handling methods is needed when harmonised cross-border datasets are disseminated. Besides looking into the legal issues related to data confidentiality, which cannot be solved by the GEOSTAT action alone, there is a need for further technical developments of disclosure control methods which take account the spatial characteristics of grid data into account.

5.3 Data Quality

GEOSTAT is convinced that, generally speaking, the quality of population grid data produced by National Statistical Institutes is higher than disaggregated European-wide population grid data.

This assumption is based on privileged access to data sources. Statistical institutes are able to use data sources of greater spatial accuracy than they are allowed to publish for general use.

There are two aspects to the notion of quality in the context of grid-based statistics:

- quality of the georeferenced source data;
- quality of the production process.

When looking at the available standards for documenting data quality with metadata, INSPIRE has a clear focus on spatial data properties such as positional accuracy and on the spatial aspects of the data production process.

On the other hand, the statistical world has developed its own standards (SDMX).[19]. The quality of grid data should be described from both perspectives, but there is currently no initiative to combine the two standards for the description of geostatistics. For the purposes of the GEOSTAT dataset the following initial list of quality characteristics and metadata properties is put forward:

Table 3: Measures concerning georeferenced source data:

Positional system	Georeferences in the form of point, polygon, or line (definition by metrer or scale)
Positional accuracy	Accuracy of georeferences
Positional source	Are the georeferences the result of a computation from official data sources or are they interpolated between known points (e.g. between addresses at

[19] e.g. the Euro SDMX Metadata Structure (ESMS), ESS Standard Quality Report Structure (ESQRS) http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/metadata/metadata_structure.

	cross roads).The 'approximately located population proportion' of the INSPIRE specification provides information about the consistency of georeferences measurements.
Logical consistency	Yes/no (no= different types of georeferences)
Usage	Known problems

Table 4: Measures concerning production of statistics by grids

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.
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 organism that takes care of them for social benefits or health insurance purposes. This is covered by the 'conventionaly located population proportion' of the INSPIRE specification.
Accuracy of the figures	In the case of figures produced by extrapolation of the sample (mainly censuses,) the measure of accuracy that results from the estimation process.
Completeness	Coverage of georeferences (% of georeferences in source data). This is covered by the 'not counted population proportion' of the INSPIRE specification.
Temporal accuracy of the spatial dimension	Last updates for the georeferences or the disaggregation sources.
Temporal accuracy of data	Last updates for the data itself (e.g. census actual data collection date).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.
Geographical coverage	Extent, total area.
Coherence	% of consistent and comparable data — regional differences in quality.
Temporal accuracy	% of same reference date.
Confidentiality	% of grid cells suppressed, thresholds for confidential data.
Quality report	Available/not available
Inspire compliant metadata	Available/not available

5.3.1 Grid sizes

The GEOSTAT project proposes that a primary grid system consists of grid cells with 10^a (where $6 \geq a \geq 0$) size in metres. The primary grids will provide a hierarchy of six scale intervals or a series of map tiles that are needed for descriptions covering the local to the global. The size of the grid cells therefore ranges from 1000 km for global studies, through 100 km for continental, down to 1 m for high resolution analysis at the suburban level. The phenomena described in a study should have a spatial resolution that is larger than the grid cell size used to provide enough level of detail. Fig 5.4 illustrates the territory that is typically covered in a reasonable way with different grid cell sizes, 1km² for the country, 100 m for the region and 10 m for the Urban Area of Stockholm.

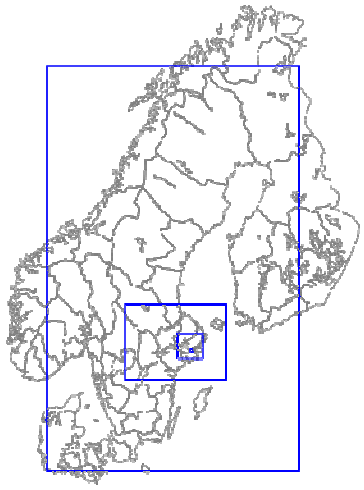


Figure 8: An example of a map tiling system applied to the administration of Sweden over a hierarchy of three administrative levels from national to commune /municipality levels (Backer 2011).

A secondary system of grids is based on a two-level quadtree solution, $10^a/2^b$ (where $6 \geq a \geq 0$ and $2 \geq b \geq 0$). This is defined for a more extensive series of grids: 10 m, 25 m, 50 m, 100 m, 250 m, 500 m, 1km. 2.5km, 5km, 10km, 25km, 50km. The advantage here is that these grids (b>0 (e.g. b=2 or 250 m grids)) will all aggregate up to the next primary grid (b=0(e.g. 1000 m grids)). The primary grid is thus a special case of the secondary grid system (Backer 2011).

6 A concept for a GEOSTAT data infrastructure

6.1 GEOSTAT and INSPIRE

One objective of this project has been to describe a Eurostat vision of a data infrastructure to access and share Geostatistics. The infrastructure work package has furthermore been dedicated to analysing the spatial data infrastructure provided by the INSPIRE project. One of the critical issues in the INSPIRE project has been its cartographic bias, although many of the data themes mentioned in the directive's annexes may be seen as integrated datasets, which are datasets based on the integration of geography and statistics. Geography and statistics are very different data types and are kept separate. However, together they amount to the two sides of spatial information implied by the INSPIRE directive. Hence, the aim of this project has been to produce a sound description of a working infrastructure for the integration of geography and statistics, giving users the opportunity to explore statistical data using a geographical interface.

The GEOSTAT dataset is a distributed service. In principle all NSIs which deliver data to the system will share the benefits of its use and will also contribute to the definition of the terms or business model under which the data may be used. Visions for a common data distribution model will be followed up in GEOSTAT 1B.

The publication of grid statistics falls under the INSPIRE directive and has to fulfil its requirements as regards metadata, interoperability and sharing as laid down in the directive and the corresponding regulations. However, being a statistical and not just a spatial dataset, there are also certain restrictions and conditions as regards compliance of the GEOSTAT dataset with INSPIRE.

This section will look at the possibilities and limitations of grid statistics for viewing and for exploration as a spatial dataset.

6.2 A concept for a GEOSTAT data infrastructure

A geostatistical data infrastructure will have to accommodate three very different data types (text, tables and graphs, and maps) and the services operated on those data. Similar to other service oriented architectures like INSPIRE or the SMDX architecture, geostatistical architecture will consist of different components.

6.2.1 Data types

For the integration and sharing of all information in a data infrastructure, data have to be broken down into reference data structures suitable for the type of information and data.

- Geosemantic information — alphanumeric texts needed for working with narratives to describe geosemantic data structures and their metadata.
- Geospatial information — geographic features used as the main component of geospatial data structures and their metadata.

- Geostatistical Information — statistical information used as the foundation for geostatistical data structures and their metadata.

6.2.2 Data services

Services will make it possible to seek, find, explore, buy and download information available in the above data types to be included in the information system tailored to a specific issue. The problem of providing unified services for geosemantic (alpha numeric) information is largely solved and can also be used to some extent in the statistical context (SDMX) - Dublin Core Metadata.

INSPIRE describes 5 service types for geospatial information:

- geospatial discovery services;
- geospatial view services;
- geospatial download services;
- geospatial transformation services;
- invoke geospatial data services.

For geostatistical information, the service types of the other two data types are not suitable. The spatial data infrastructure (SDI) for geostatistics will require new types of services that have to be able to operate on existing infrastructures (SDMX, INSPIRE) to make the full range of information available in an integrated manner:

- geostatistical discovery service, most probably using SDMX and components such as the SDMX repository:
- geostatistical view services. There have been discussions in connection with view services spatially adapted to geostatistical information. The minimum required is:
    - geostatistical map service. The map service developed for the INSPIRE project is based on the use of web map services (or web feature services). In connection with view services for geostatistical information, it could be argued that web processing services would be advantageous (specially adapted to maps with grid data) because this service would enable users to select any part of a grid dataset and obtain table and diagram presentations of the result.
    - geostatistical table service. A table service would make it possible to generate a well structured table from a given dataset selected with a web processing service.
    - geostatistical diagram service. A diagram service would make it possible to generate a diagram from a dataset selected with a web processing service.
- Geostatistical download services. In addition to the technical aspects, download services will provide a solution to the business side of a ‘rights management system’.
- Geostatistical business service. A business service would make it possible to handle questions about rights, prices and conditions that would also facilitate transactions in cross border situations.
- Geostatistical transformation service.

- Invoke geostatistical spatial data services

6.2.3 Rights management

When discussing a spatial data infrastructure (SDI) for a GGI (see section 0), it is essential to develop a sustainable solution for 'Rights management' as well as a 'Business model' for all three data types.

'Rights management' is one of the key issues that must be settled if a geostatistical information system is to work. While most, if not all, official statistical data from public sources in Europe is normally free of charge and can be used provided the source is acknowledged, this is often not the case with geostatistical information. There is currently a great deal of research going on in the spatial community on digital rights management and the development of authorisation and access control services. In the context of this project the problem is mainly how to make geostatistical high resolution data available as easily as possible, preferably free of charge. In most cases this is a decision to be taken by the management of an NSI, the ESS or lawmakers at national or European level.

6.2.4 Metadata for GEOSTAT data

Publication of metadata on the content of the grid dataset is not as straightforward using ISO 19115. Metadata on a grid dataset have to describe multiple statistics referenced spatially to one set of geometry. Publication of metadata which is compliant with INSPIRE/ISO 19115 is primarily about information of a dataset on one single feature type (i.e. land cover or buildings).

Metadata on statistics have to incorporate metadata on many different statistics for one specific spatial delineation. More specifically, in the case of grid statistics many statistics are published for one specific grid net or even grid cell.

Hence, simply adopting a metadata standard on grid statistics from the INSPIRE Annex III theme 'Population distribution' will not work, because metadata on many other statistics will have to be created as well (buildings, land use, tree densities etc.). The Thematic Working Group on 'Population distribution' and 'Statistical Units' has therefore suggested additional metadata on the level of feature type and data type.

According to these recommendations metadata on statistics for grids will have to be published at three levels:

1. Dataset: Refers to a set of statistics for one specific spatial delineation — INSPIRE Metadata will be published. A reference to the content of different statistics has to be made in the field 'Lineage'.
2. Statistical set: Refers to a specific statistical set — Each specific statistic will be described by metadata. It has not been decided where these XML-files will be published centrally. For the moment they will be stored together with the data.
3. Data values: Refers to individual values within a statistical set — Contains a reference to a grid cell and metadata on each of the specified values.

Table 5 contains an example of metadata of a statistical set and metadata at the level of data value. This example describes:

- Total number of men less than 25 years old on 1 January 2010 per 1kmx1km grid in ETRS89-LAEA projection for the Netherlands.

About 4000 males are not located at an individual address. Individual values and their metadata are given for a regular and a disclosed grid cell.

Table 5: Example of metadata on Statistical Data Distribution. Netherlands.

Name	**Data type/Code list**			**Value**
Statistical set				
Inspire Id	Identifier (codelist)			code? (to be defined)
Area of dissemination	Statistical Unit (total AREA!!):			Netherlands, statisticalGrid
Classifications	Classification 1 (codelist)			Sex
Domain	Domain (codelist)			Demography
Measure	Variable (codelist)			Population at place of residence
Measurementunit	Unit of measurement (codelist)			Persons
Not-counted population proportion	(number)			4000
Period of measurement	TM period (date)			15-02-2010
Period of reference	TM period (date)			01-01-2010
Period of validity	TM period (date)			-
Statistic measurement method	Statistics measurement method (codelist)			Count
Universe	Universe (text)			Under 25 years (as a subset of classification)
Data value	**Statistical Data Value (individual, regular)**			
		Approximately located population proportion (number)		1
		Comment (text)		Partly redistributed people, contains homeless people
		Conventionally located proportion (number)		4
		Dimensions: Statistical Data Dimensions		
			Classification 2 (codelist)	Male
			Georeference: Statistical unit (codelist)	Grid (ETRS89-LAEA): 1kmN3079E4032 (Statistical l GridCellUnit)
		Flags		-
		SpecialValue* (codelist)		-
		Value*		16 (total?)
	Statistical datavalue (individual,undisclosed)			
		Approximately located population proportion (number)		-
		Comment (text)		-
		Conventionally located proportion (number)		-
		Dimensions: statistical data dimensions		
			Classification2 (codelist)	Male
			Georeference: Statistical unit (codelist)	Grid (ETRS89-LAEA): 1kmN3080E4033 (StatisticslGridCellUnit)
		Flags		
		Special Value* (codelist)		Confidential
		Value*		-

* Either Special Value or Value must be provided.

One of the issues for the statistical set missing here seems to be the status of the data. This status should be included in a metadata field ‘Status’ and encoded using a code list (‘preliminary', 'definitive’, etc.).

The specifications in the Annex, Theme III, "Population distribution" are not yet final. When applying the draft guidelines, the publication of metadata is foreseen at different levels:

- Publication of value metadata will be at the level of the individual data.
- Publication of metadata on the statistical set will include the set of values.
- Publication of metadata on the grid dataset will be according to INSPIRE metadata. A link to the dataset is part of this metadata.

Metadata on the grid dataset will be published for each country on the national geoportal.

6.2.5 Model for a Statistical Web Service

Statistical Web Services will be used to implement the GGI and connect its various information systems. The core requirements for such a service include: the selection of the spatial area of interest; the statistical unit and variable; output of a map covering the area; a table or set of tables with appropriate statistics, and a diagram showing the distributions of variables.

A prototype web processing service has been developed to demonstrate, as a proof of concept, what a statistical web, designed for interactive on-screen data selection and obtaining statistics from it, might look like.

The user selects an area by means of a buffer around a point. The radius of the buffer is modelled as a parameter to be altered by the user. The user selects one or more points on screen and may alter the radius of the area. Then, aggregated statistics on the subject "Total population per 1km² grid' are presented on screen (see example in Figure 9). This little study already integrates the basic concepts of the user interface and back-office processing of a statistical data infrastructure such as statistical data selection, spatial selection of the area of interest, on-line processing of the statistical parameter in response to the selected area and presentation of the information to the user on screen. The combination of database type processing and geoprocessing will allow various interactive applications to serve the basic publish-find-bind pattern of open service-oriented data infrastructure.

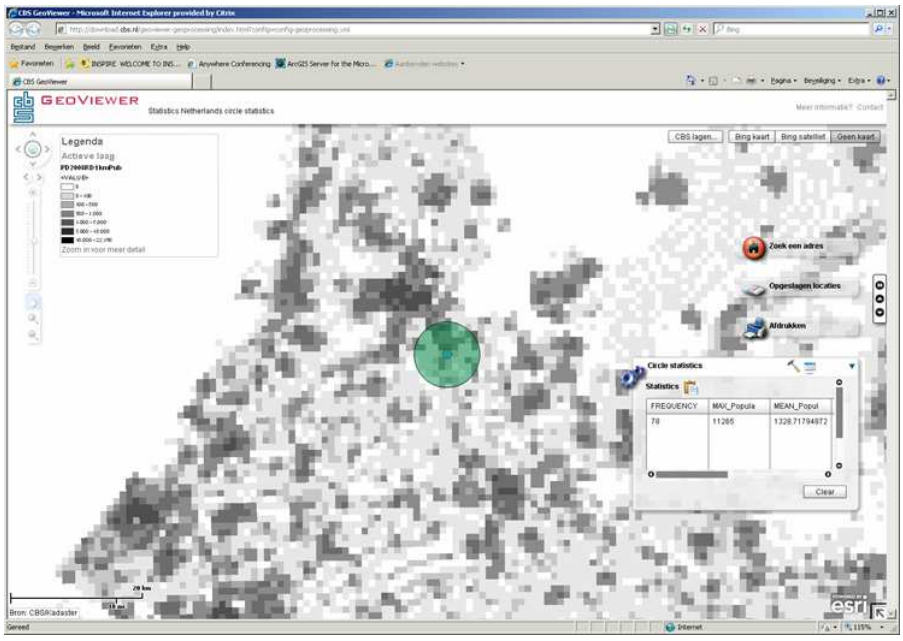


Figure 9: Geoprocessing web service

7 Conclusions and further work

These conclusions are based on surveys, numerous discussions with a broad audience within the ESS and results gained during the project. Some issues are not yet settled and need to be worked further within the ESSnet GEOSTAT 1B project.

User needs have been collected on the basis of many years' experience in the countries which provide data on grids and where the grid data are used for various issues of analysis by the statistical offices themselves, other public authorities and the commercial market. Furthermore, the preconditions and possibilities for disaggregating data in the non-grid countries have been investigated.

The collected cases have been examined with respect to type of analysis, use of grid data and constraints of data interoperability experienced. Furthermore, the action has developed the data specification, suggesting solutions for core concerns such as map projections, the question of scales, the coding system, methods for spatial analysis, methods for delineations and confidentiality issues.

This action has improved the overall scope and quality of the existing 'ESSgrid' effort to map the European population on 1 km² grids. In previous work, national data sets did not contain harmonised data and showed one variable only, that of total population.

The project has described and tested disaggregation algorithms, INSPIRE specifications and data protection rules for population grid data (both disaggregated and aggregated). The project has furthermore described and tested aggregation methods, including housing and agricultural holdings grid data. A first set of methods to produce harmonised grid data has been discussed with Eurostat and with all the participants. Examples of harmonised grid maps and statistics from the whole of Europe as well as from specific parts of Europe have been produced by other participating countries. The project has visualised harmonised European population grid data and demonstrated the use of grid data in spatial analysis.

7.1 Knowledge sharing

The objective of the action has been to spread the knowledge and results gained in the project to the ESS. The EFGS functions as a professional reference group for the GEOSTAT 1A ESSnet project. This network of experts, comprising key users and producers, has scrutinised the output of each iteration and suggested improvements to the project results. As the results in the first instance concern statistical offices, one focus of the knowledge transfer has been the annual meeting of the GISCO working party with the NSIs and the national mapping agencies.

The action has made sure that non-participating countries benefit from the experience and are encouraged to adopt and implement the approach. In March 2012 a course within the ESTP framework is being set up and will build upon experiences in GEOSTAT 1A. The final report will also be presented at the annual GISCO WP in March 2012. The EFGS website (www.efgs.info) has been used for all digital publishing.

7.2 Sustainability

The ESS partnership builds upon the existing EFGS network and website. During the GEOSTAT 1A project several NSIs have contributed on a voluntary basis, and many NSIs have nominated a national contact person. Today the EFGS network consists of national contacts from 32 countries.

To make the action sustainable, support from the ESS has been requested, among other things to solve organisational issues such as common rules for handling confidentiality and to motivate more countries to produce national grid datasets using national data sources.

Special attention is required to ensure that grid datasets can be produced not only from census data every 10 years but that the datasources are updated continuously and will allow the production of grid datasets at shorter intervals.

7.3 Grid production

At this stage the project has concentrated on total population. Ongoing development of address files and geocoding of buildings together with the census data of the year 2011 will give more opportunities to extend grid data variables or the scale of the geographical grid cells. Regarding grid sizes, the GEOSTAT project proposes *a primary grid system* consisting of 10^am (where $6 \geq a \geq 0$) grids. The primary grids will provide a hierarchy of 6 scale intervals or a series of map tiles that are needed for descriptions ranging from the local to the global.

We also suggest introducing intermediate grid sizes, *a secondary system of grids* based on a two-level quadtree solution, $10^a/2^b$ (where $6 \geq a \geq 0$ and $2 \geq b \geq 0$). This is defined by a more extensive series of grids; 10 m, 25 m, 50 m, 100 m, 250 m, 500 m, 1km, 2.5km, 5km, 10km, 25km, 50km. The advantage here is that these grids (b>0 (e.g. b=2 or 250 m grids)) will all aggregate up to the next primary grid (b=0(e.g. 1000 m grids)).

Conversions of grid data were tested with different sizes of national grids. Conversions were made by using the centre points of grid cells as reference points. If a conversion from one projection system to another is needed, it is better to do it on the source data.

Support for the development and refinement of disaggregation methods in cooperation with research institutes will remain high on GEOSTA's agenda, as this is the only way to guarantee updated and more diversified grid datasets of Europe.

7.4 One package — one provider

Users' experiences show that accessing a country's grid-based data is not always easy. Therefore ‘one package — one provider’ is the preferable way to obtain data conveniently and quickly. The goal of the EFGS is to act as a hub for users of grid statistics. It should act as a single point of contact for users and provide links to data (with or without small restrictions) as well as links to the contact points of providers who supply data with restrictions. *Business models* differ from one country to another. Some data are distributed free of charge; others request a fee for delivering data to help produce the statistics. With a view to establishing ‘one package — one provider’, a comprehensive business model should be developed in GEOSTAT 1B. The ultimate goal remains to distribute the GEOSTAT dataset free of costs and without usage restrictions in the same way as statistics in the ESS.

The EFGS will not sell data involving statistics of different variables until a business model and the organisational aspects of EFGS have been decided upon. For data in addition to the GEOSTAT 2006 dataset, customers will still have to contact each national contact person. We will, however, make a suggestion for a common business model in the follow up.

In most countries data are published in a national grid system based on their national reference system. This has the advantage of fitting into their national geolayers and producing square grids. However, this conflicts with the wish for a harmonised system that covers continental Europe. On a European level the Grid_ETRS89-LAEA, as defined by INSPIRE, is now accepted and GEOSTAT has therefore decided to collect data in the 1km grid of that system.

Respondents from both the producers' and users' sides rated grid sizes up to 1km² highest. For grids larger than 1 km² importance declines with the increase of grid cell sizes. Data users who wanted the grids for areas of a country and smaller prefer the smaller grid sizes up to 250 m², while those who required the data for groups of countries or continents preferred the 1km² grids.

Users mainly need population, housing and economic variables on the basis of grids. This corresponds to availability in most grid-countries. Data such as total population figures, number of buildings and dwellings are generally available on the basis of grids for grid sizes of 1km² or even smaller. Generally speaking, however, the data is released only under certain conditions, such as agreeing to permission agreements, payments having to be made and referencing to NSIs being required.

Data providers should already be looking towards the period after the 2011 census and develop an update policy for the grid datasets. An annual dataset is the preferred option.

7.5 Confidentiality

As a general rule, it can be concluded that data providers differentiate between more and less sensitive data and define accordingly the thresholds for variables and corresponding grid cell sizes.

One of the objectives of this ESSNet project GEOSTAT is to create a 1km² population grid map and dataset of Europe. As the 1km² grid can be considered to be a fairly rough grid, and as we are only asking for absolute numbers without further socio-economic variables attached, we offer the following solution for the harmonised European population grid. No restrictions will be applied to the publication of total numbers such as the total population figure, the total number of buildings and the total number of dwellings. *Confidentiality* problems become more obvious when the amount of variables to be gridded is extended or when data are delivered by grid sizes smaller than 1 km^2. An agreement is required on the use of similar confidentiality handling methods when harmonised cross-border datasets are disseminated. Furthermore, there is a need to develop disclosure control methods which take into account the spatial characteristics of grid data.

Future work will include finding a solution for further variables, in particular those which break down these total numbers (e.g. population by age group and sex...). The project recommends that data *protection measures* should depend on the sensitivity of the variables. The absolute count of the respective units (such as number of inhabitants, households, buildings, workplaces) should be disseminated without any restriction, even for the smallest grid sizes.

One way to control dissemination and the use of grid-based statistics is to release them only when a licence to use is granted. There is an urgent need for further development and harmonisation of disclosure control methods for spatial statistics by grids.

7.6 Quality

The GEOSTAT project has examined a common process of *quality assessment*. Quality assessment has to include different approaches (geographic, statistical, production, standards, etc.) The assessment may also differ according to the data sources.

D*isaggregation methods* are used, for example, when global or European-wide population grids are made. The choice of source data is dependent on data availability and the size of the grids in the final output. Spatial statistics by census enumeration areas are usually the best starting point. Validation of the results against the area-based source data is an important part of the process and may lead to a refinement of the original model.

Statistical institutes are able to use data sources which are more spatial accurate than those they publish for general use. However, if grid-based data are nationally available they are produced by quite diversified methods and by different types of data sources. Documentation of the quality of data sources tends to be quite poor and production methods are just developing.

In the area of *grid-based statistics* there are two major perspectives: one is the quality of its georeferenced source data and the other is the quality of the production process. INSPIRE mainly standardises metadata descriptions and quality measures concerning georeferences and spatiality, focusing on elementary data accuracy for these topics. But the statistical world has its own standards (SDMX). The quality of grid data should be described from both perspectives.

7.7 Dissemination

The project has disseminated population grid data by 1 km^2 free of charge via the EFGS web site in the form of text-files (*.csv-files). The role of the EFGS as a data provider and the definition of a common dissemination policy, at least for some grid-data deliveries, are the issues that need to be agreed on in the future. The role of the INSPIRE model of decentralised data delivery via open interface must be discussed as well.

7.8 Tools

For the production of future iterations of the GEOSTAT dataset at both the European and the national levels, more support for tools is needed and the tools should be harmonised. To allow for use in all IT environments tools should be non proprietary and open source.

8 References

Backer, L., Tammilehto-Luode, M., Gublin, P. (2002). Tandem_GIS_I. A Feasibility study towards a common geographical base for statistics across the European Union. Eurostat. Working papers. http://europa.eu.int/comm/eurostat/Public/

Backer, Lars (2012). In search of an SDI for Geo Statistical Information. www.efgs.info

Backer, L. , van Leeuwen, N., Bloch, V., Kaminger, I., Tammilehto-Luode, M. (2011). In Search of a Hierarchy and Coding System for the Grid ERTS89-LAEA. www.efgs.info

Backer, Lars, H. (2011). Towards a SDI for Geo Statistical Information. European Forum for Geostatistics.

Batista e Silva, F., Gallego, J., Lavalle, C. (2011). The effect of ancillary data in population dasymetric mapping. European Forum for Geostatistics, Lisbon, October 2011.

Bloch, Vilni Verner Holst. Gundersen, Geir and Thorsdalen, Bjørn (2010). A State of Art -report Norway. v2. 2010-09-09. 12 pages. www.efgs.info

Corcoran, Dermot (2011). A population grid for the Republic of Ireland: Making use of national databases and local geography. European Forum for Geostatistics, Lisbon, October 2011.

D2.8.I.2. INSPIRE Specification on Geographical Grid Systems — Guidelines, version 3.0

D2.8.III.1. Data Specifications on Statistical Units — Draft Guidelines, version 2.0 (2011_06_15).

D2.8.III.10. Data Specifications on Population Distribution — Draft Guidelines, version 2.01 (2011_07_13).

EFGS (2010). A State of the Art report Summary v1. 2010-09-24. 17 pages. www.efgs.info

European reference grid 2003. 1st Workshop on European Reference Grids. Short Proceedings. Ispra 27-29 October 2003. JRC ESDI Action 2142.

Eurostat (2011). Proposal for new explanatory notes for degree of urbanisation. Working group Labour Market Statistics. 6-7 June 2011. Doc.: Eurostat/F2/LAMA/05/11

ESS Eurogrid Population Map 2009, http://www.efgs.info/presentations

Gallego, Javier (2010). A population density grid of the European Union. Population and Environment. Springer Science+ Business Media, LLC.

Goerlich, Fransisco and Cantarino, I. (2011). Downscaling Population with a High Resolution Land Cover Data Set for Spain. EFGS Conference. Lisbon, October 2011.

INSPIRE metadata editor: http://www.inspire-geoportal.eu/index.cfm/pageid/342

INSPIRE Specifications on Geographical Grid System — Guidelines. 2010. D2.8.I.2. (http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_Specification_GGS_v3.0.1.pdf)

Jablonski, Radoslaw (2011). A State of the Art report Poland. v2. 2011-11-21. 19 pages. www.efgs.info

Kaminger, Ingrid (2010). A State of the Art report Austria v1.0 . 2010-09-24. 9 pages. www.efgs.info

Kaminger, Ingrid (2011). Disaggregation of population densities. Ways out of the CORINE deadlock. GISCO WP. Eurostat. Luxemburg. March 2011.

Koivula, M. Tammilehto-Luode, Marja. Tammisto, Rina (2011). A study of conversions from national spatial data to harmonised European grid data by Statistics Finland, http://www.efgs.info/data/geostat/geostat-methods

Lipatz, Jean-Luc (2011). The French toolbox. Generating population gridded data by aggregation — simple tools. Appendix 2.

Lipatz, Jean-Luc (2010). A State of the Art report France v1.1._f. 2010-09-13. 20 pages.. www.efgs.info

Lipatz, Jean-Luc (2010). Gridded population data by INSEE. European Forum for Geostatistics, Lisbon, October 2011.

Lipatz, Jean-Luc (2010). Gridded data from the French census 2007. Aggregation without coordinates, coordinates but disaggregation. JL. Lipatz 23/11/2011.

Masik, Kreet (2011). Production of grid-based statistics in Statistics Estonia. European Forum for Geostatistics, Lisbon, October 2011

Santos Ana Maria (2010). A State of the Art report Portugal. v2. 2010-09-22. 19 pages. www.efgs.info

Schoenmakers, B-J. 2011. Creating 2001 to 2011 population grids using Census Geography. Statistics Portugal. European Forum for Geostatistics, Lisbon, October 2011.

Sehlin, Johnny (2011). Production of grid statistics at Statistics Sweden. European Forum for Geostatistics, Lisbon, October 2011.

Sommer, Erik (2007). Presentation of Danish method to take care of confidentiality when providing statistical data on grids. Nordic Forum for Geostatistics. Helsinki, Finland, September 19-21 2007.

Statistics Portugal (2011). Portuguese Population Grids for 2001, 2006 and 2011. Paper on the Methodology.

Steinnocher, K., Kaminger, I., Weichselbaum, J. Köstl,M. 2010. Gridded population –new data sets for an improved disaggregation approach. European Forum for Geostatistics. Tallin, October 2010.

Tammilehto-Luode, Marja (2010). A State of Art -report Finland. v2. 2010-08-30. 15 pages. www.efgs.info

9 Annexes

9.1 ANNEX I: User Needs Survey

As part of the GEOSTAT Project, the European Forum of Geostatistics (www.efgs.info) conducted a survey to learn about users' needs in respect to grid based statistics. Analysing users' needs in relation to the grid based statistics is one of the main aims of the GEOSTAT project.

The survey took place in September/October 2010. It was sent to the members of the NSI and NMA working group who participated in the annual Eurostat working group for GIS for Statistics, and also to grid customers. Preliminary results were presented at the EFGS-Conference in Tallinn in early October. Further answers were submitted after that and are included here. Between 26 September 2010 and the deadline of 28 October 2010, when the survey was closed, 45 answers were submitted; 37 of these were fully completed.

The survey included voluntary fields for contact details, but it obviously could also be completed anonymously. Apart from 5 anonymous answers, 8 answers came from data users from Austria, 5 from Finland, 3 each from Denmark and European institutions, 2 each from Estonia, France, Norway, Slovakia, Spain, Sweden and the UK and 1 each from Bulgaria, Croatia, Czech Republic, Germany, Ireland, Malta and the Netherlands.

About 30% of the respondents work in governmental institutions, 20% each in Research and for private organisations and 16% in education. The survey respondents work in a wide range of fields. Amongst the answers given the most popular activities were environment, regional and local planning, infrastructure/transportation and marketing. A number of respondents noted that their work activities were not defined in the survey. For example, some of them work in fields such as education/research, economy and health.

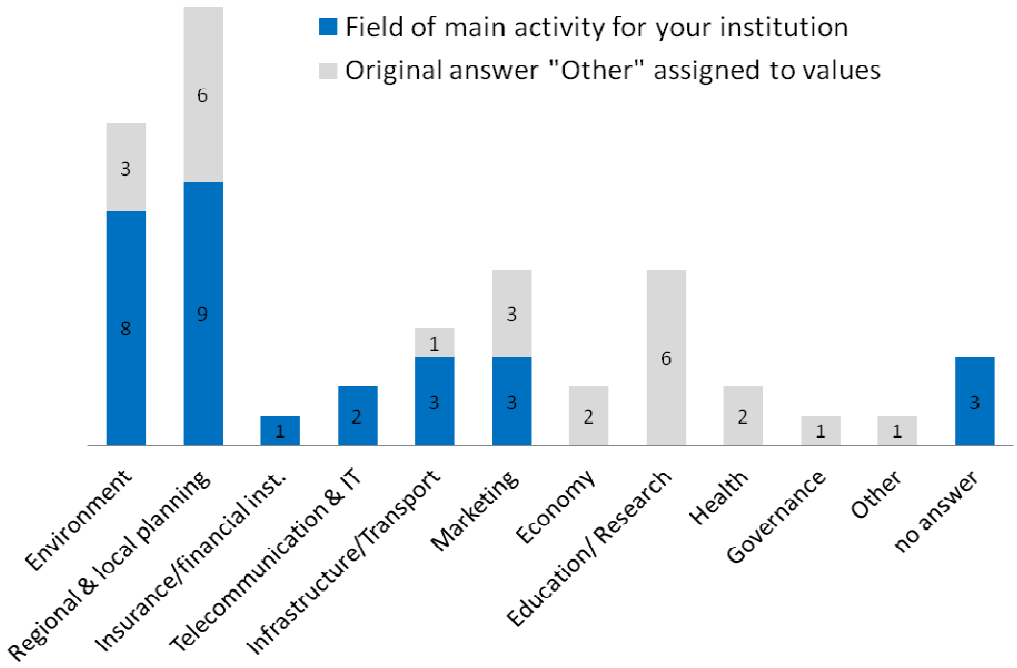


Figure 10: Field of main activity of respondents' institutions

The majority of the respondents use grid based statistics as an input in different spatial analysis tasks. The most frequent subjects for spatial analyses are regional and local planning, demography and environment. Some other minor, but nonetheless interesting subjects are business, marketing, health and real estate, as well as heat and energy demand.

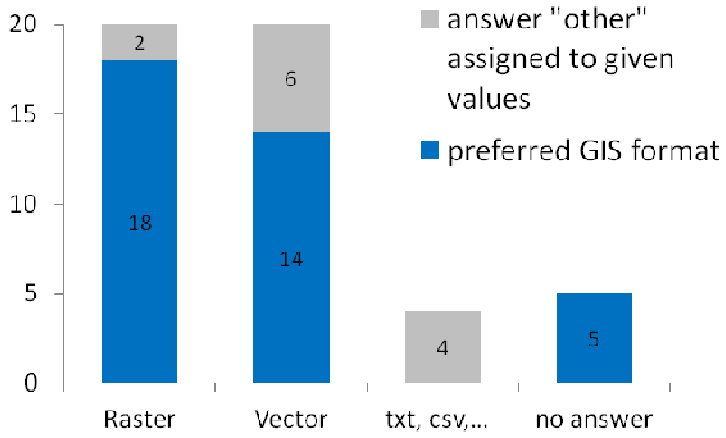


Figure 11: Preferred GIS format to receive grid based data in.

GIS compatible raster and vector formats are the most preferred. Some respondents also noted the need for txt and csv formats. It can be presumed that they understand this as separate delivery of geometry and data with unique identifiers, which can then be used to connect the data to the geometry. In this way, the data volume of transmitted or downloaded files can be reduced significantly.

For the survey respondents the most important grid cell size is the 1km grid. In fact, there is a clear tendency on the part of most respondents to regard all resolutions from 100m x 100m up to 1km x 1km as very important or important. For grids larger than 1 km, the importance declines as grid cell sizes increase. Up to grid sizes of 10km they are still considered fairly important, whereas grids larger than 10km are mostly seen as unimportant.

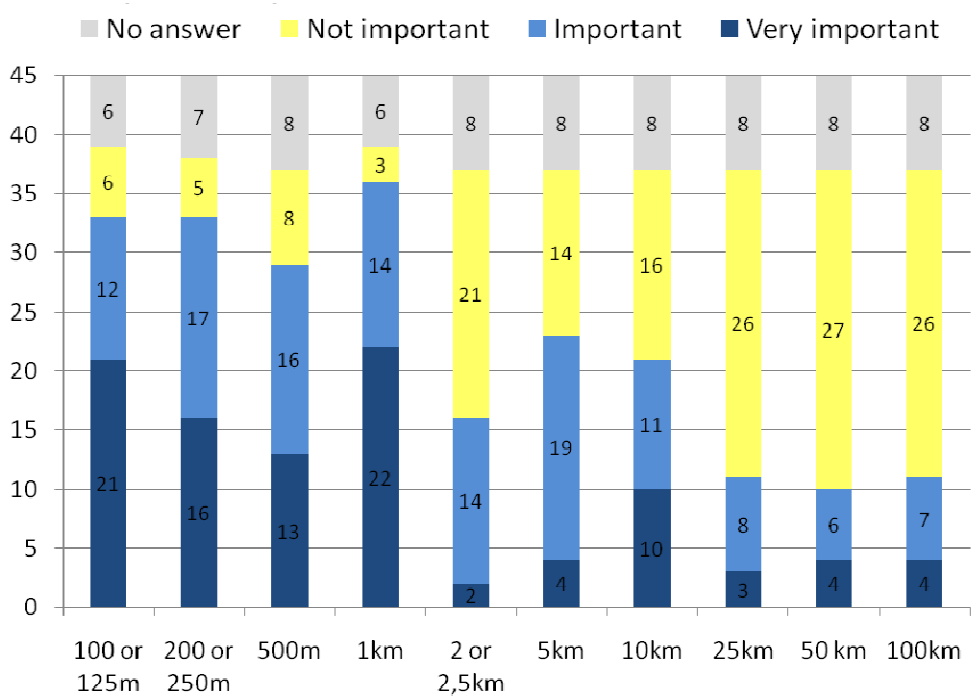


Figure 12: Importance of various grid sizes

There was also a clear correlation between area of interest and grid size. Data users who wanted the grids for areas of a country and smaller preferred the smaller grid sizes up to 250m, while those needing the data for groups of countries or continents preferred the 1km grids.

As grid data may not be disclosed without restrictions for data confidentiality reasons, the respondents were asked their opinion on the different methods of data disclosure. All methods obviously have one of the two disadvantages – either loss in data or loss in spatial accuracy. The two most frequent methods in use were given as options, including a brief explanation of what was meant.

- Aggregating grids. The layer of the data contains grids of different size filling up the whole territory.
- Replacement of undisclosed values with another number or character (for example 99999999). This method will result in loss of data, but the grid size remains the same throughout the dataset.

The most preferred method for data disclosure (> 50%) is the replacement of undisclosed values by another number or character. Nevertheless some respondents (20%) prefer the grids with undisclosed values to be aggregated with other grids and a further 20% take the view that completely different methods should be used for data disclosure.

There was no option for specifying what “other” meant. However, the comments at the end of the survey suggest that there is indeed a need to disclose absolute numbers of Population and Buildings.

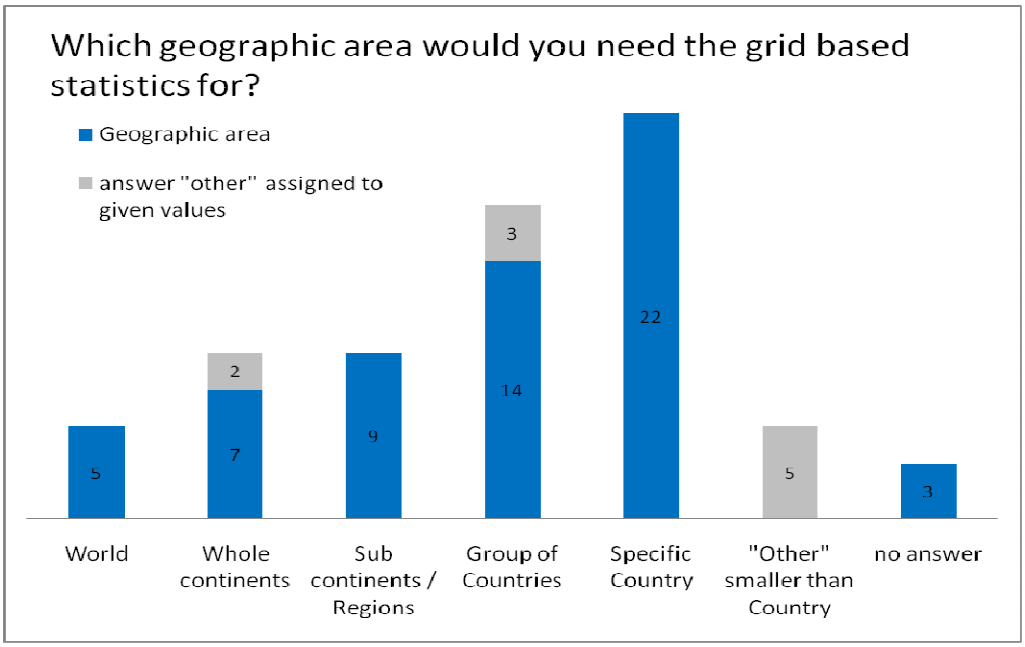


Figure 13: Geographic area the grid based statistics are needed for

As Figure 13: Geographic area the grid based statistics are needed for
Shows, the highest demand for grid based statistics concerns specific countries (49%). 38% of the respondents need grid based statistics for a group of countries and 20% each for Sub-continents/Regions and whole continents. A world perspective is not particularly useful for European users at the present time. Multiple answers were allowed.

A further question simply inquired about the needs of users with regard to sourcing their data. Would respondents like to receive the grid based datasets covering the whole of Europe in a single package from one provider (for example from EFGS)? Currently, the grid based statistics can be ordered from some national statistical institutions. The spatial territory is normally a country and/or a part of a country.

A clear majority (>70%) ticked the option “Yes”, which indicates that, for users, the one package/one provider option is the preferred option to obtain data conveniently and time efficiently. The option “No” was mainly chosen by customers interested in grids on a local level who are involved in regional planning and analysis, mostly within a particular country.

The concluding questions inquired about the fields of statistics for which the respondents would need data on the basis of grids. Multiple answers were possible. According to expectations, the greatest need is for statistics on grid based population (84%), housing (67%) and economy (67%). About half of the respondents need statistical grid data on the environment and labour force.

The final box for further comments was mostly used to confirm the need for grid data sets in general, with the emphasis on harmonisation.

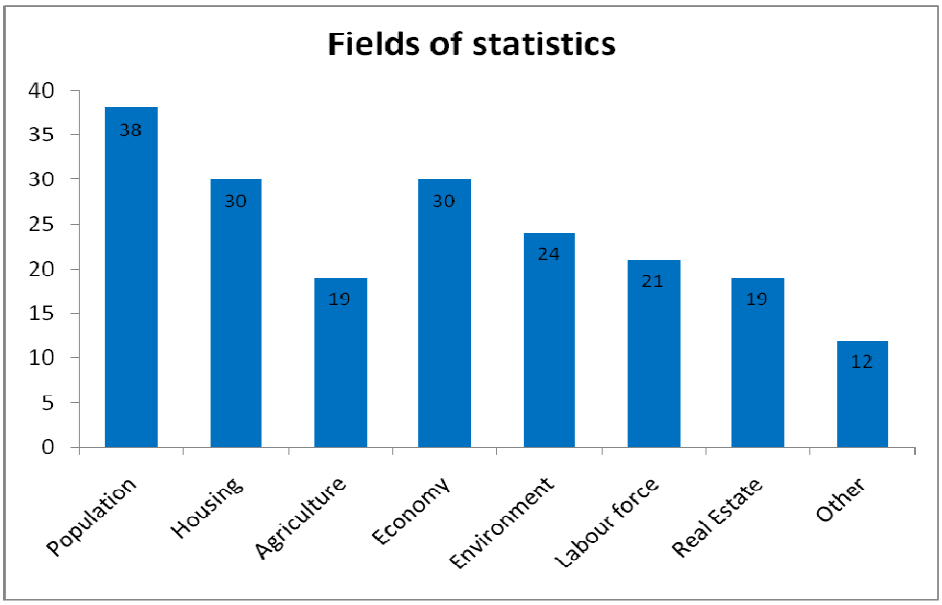


Figure 14: Fields with need for grid based statistics.

9.2 ANNEX II: Data Providers' Survey

Between 15 November and 1 December 2010 the European Forum of Geostatistics (EFGS) carried out a survey to analyse the availability of grid based statistics and related conditions of dissemination in Europe.

The survey was sent to the EFGS-contact persons and EFGS-conference participants, the NSIs, the European Census leaders and other known grid data providers both within Europe and outside. To reach further possible grid data providers, the contact persons were kindly requested to forward this survey to the person responsible for the Census as well as any organization, institute or company producing grid based statistics. By the deadline, 19 responses were received from various institutions of the countries of Austria, Belgium, Cyprus, Denmark, Estonia, Finland, Ireland, Italy, Netherlands, Norway, Poland, Republic of Kosovo, Serbia, Slovak Republic, Slovenia, Sweden and the USA. A further four responses were received by email from institutions in Italy, Malta, Turkey and Spain, explaining that they were not yet able to provide answers to the survey.

The vast majority of respondents (84%) are either grid data providers or have plans to establish a grid in the near future. Another 10% are in the process of preparing grid based statistics. Italy alone pointed out that they have no plans for grid based statistics yet.

9.2.1 Projection systems

Most countries use some sort of national Transverse Mercator Projection and base their grid system on it. Austria and Belgium mention the INSPIRE grid system based on the Lambert Azimuthal Equal Area Projection. The EPSG-Codes can be found under http://www.epsg-registry.org or http://spatialreference.org/.

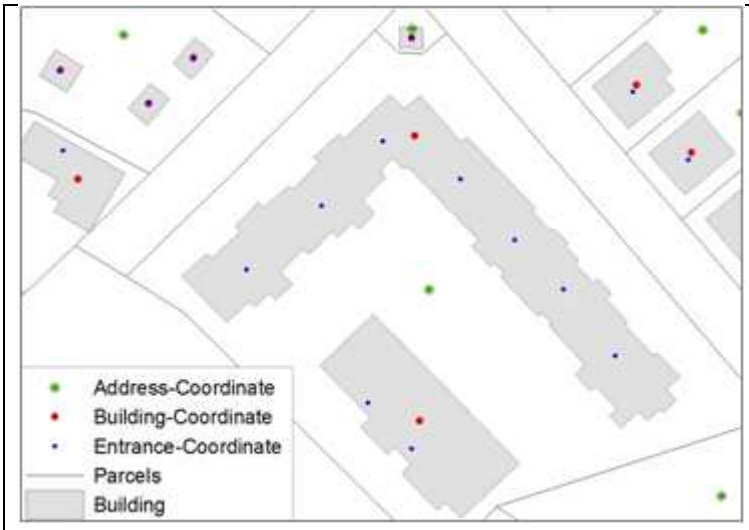
Table 6: Institutions and the projections their grid systems are based on

Country	Name of institution	Grid system	Projection and coordinate system (EPSG –Code)
Austria	AREC Raumberg-Gumpenstein	Yes	EPSG:3035 (ETRS89 - LAEA)
Austria	Statistik Austria	Yes	EPSG:3035 (ETRS89 - LAEA)
Belgium	Statistics Belgium	Yes	EPSG:3035 (ETRS89 - LAEA)
Cyprus	Statistical Service of Cyprus	Yes	CGRS_1993_Cyprus_LTM
Denmark	Statistics Denmark	Yes	UTM 32N, datum EUREF89
Estonia	Statistics Estonia	Yes	EPSG:3301
Finland	Statistics Finland	Yes	EPSG:2393, EPSG:3047
Ireland	Central Statistics Office	No	
Ireland	Central Statistics Office	Yes	EPSG:29900
Netherlands	Statistics Netherlands	Yes	EPSG:28992
Norway	Statistics Norway	Yes	EPSG:32632
Poland	Central Statistical Office	Yes	

Republic of Kosovo	Statistical Office of Kosovo	Yes	UTM WGS 1984 NH
Serbia	Statistical office of the Republic of Serbia	No	EPSG:3046
Slovak republic	Statistical Office of the Slovak Republic	Yes	EPSG:4258
Slovenia	Statistical Office of the Republic of Slovenia	Yes	GK/48 & ETRS/TM96
Sweden	Statistics Sweden	Yes	mainly EPSG:3021 and EPSG:3006
USA	CIESIN, Columbia University	Yes	Geographic

9.2.2 Geo-referenced data

To find out about the most precise positional accuracy saved with the institutions base data, there were seven options of accuracy to choose from. 71% of the respondents have geo-coordinates for each building or even more precise specifications (one per entrance), which means that for these institutions the bottom-up approach to aggregate grids is possible. In fact, the above plus two further respondents (79%) already use or are planning to use the Bottom-up approach to aggregate geo-referenced data to grids.



Spatial accuracy in survey respondents	#
Geo-coordinates for each building-part (one per entrance)	7
Geo-coordinates for each building	6
Geo-coordinates for each land parcel address	0
Geo-coordinate for a street section	0
Census / Enumerator area	3
Municipality	0
Other	2

Figure 15: Different positional accuracies for coordinates and replies to the Question: What is the most precise spatial accuracy saved with your base data?

The cell naming methods in use are mostly deduced from the coordinate values of the grid cell. 50% of the respondents, mainly the institutions of grid countries (those represented on the European grid map 2010) deduce the name of the grid cell from the coordinate values (of a corner or the midpoint). Only Estonia names them by consecutive numbers, Finland uses both methods and 7 respondents did not know.

9.2.3 Availability and rating of grid sizes

The aim of the following question was to find out two things, namely which grid sizes are available or are due to be offered, and a rating of how important/relevant/useful they seem to the respondents in their experience.

However, due to the restrictions of the PLONE programme used to conduct the survey, the radio buttons allowed only one answer per cell size. With hindsight, this question should have been split into two separate questions. One cannot necessarily deduce from the availability that a grid size is considered to be important, nor can one deduce from a rating of importance that it is actually available. Hence it was the author's interpretation to combine these answers with answers to other questions in order to check availability.

This showed that the majority of countries either already provide or will be providing data on the basis of 1km grids, and that where geo-referenced data exists and the bottom-up method is used, smaller grid sizes from 100m and 125m respectively are in use. Some countries state that they do not have grid sizes larger than 1km available, but in all cases except one these can be constructed by aggregating smaller available grids.

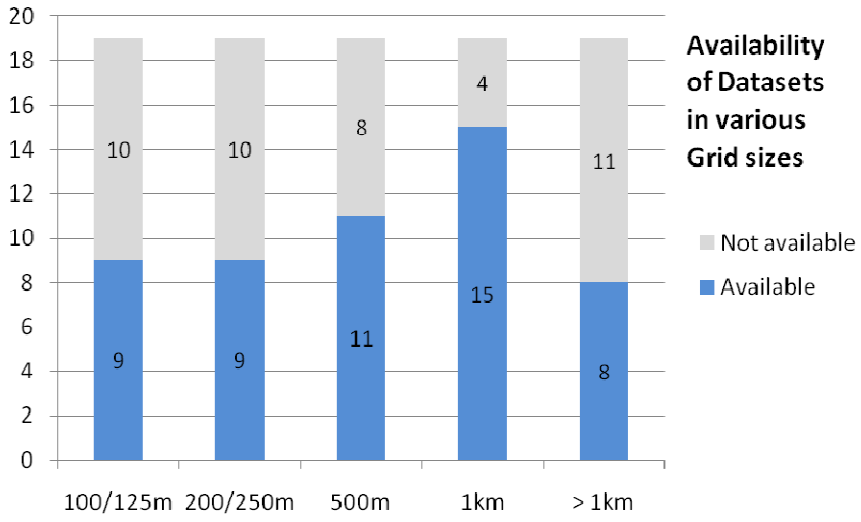


Figure 16: Availability of Datasets in various grid sizes.

Although the abovementioned argument that one cannot deduce from the availability that a particular grid size is necessarily very important, the fact that it is available gives it a 'raison d'être' and can also be included in the answers to the rating of the grid size. Two charts are set out below, one including answers and the other omitting them. Both charts show that grid sizes up to 1km are rated highest. In order to make the chart clearer, some grid-sizes are combined. The fact that two respondents consider the 500m grid to be "not important" might be seen as an outlier in the first chart, but this can be explained. These respondents have smaller grid sizes available in addition to 1km, and therefore do not need an intermediate grid size.

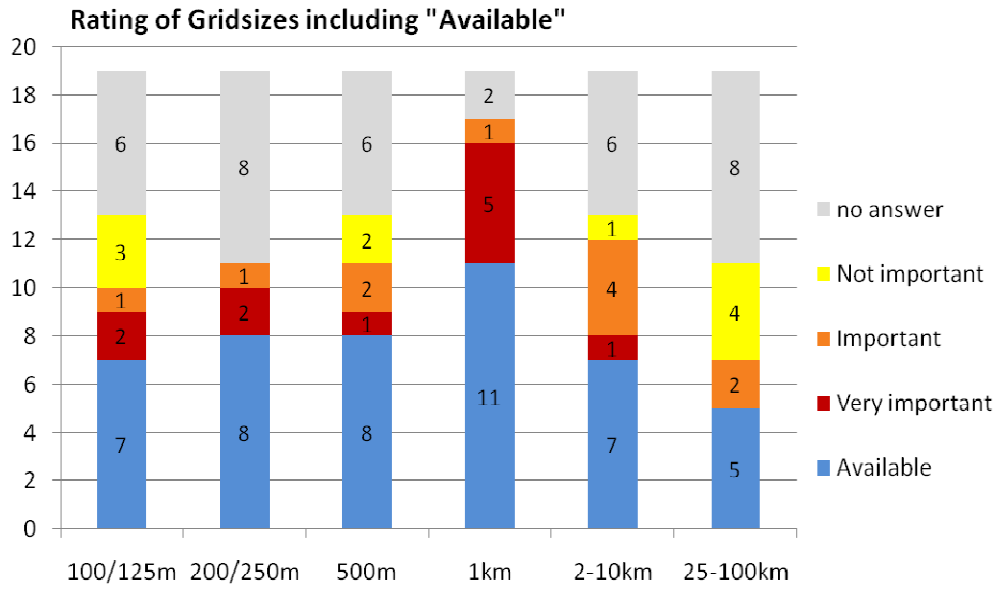


Figure 17: Rating of grid sizes including the option "Available" as "important enough"

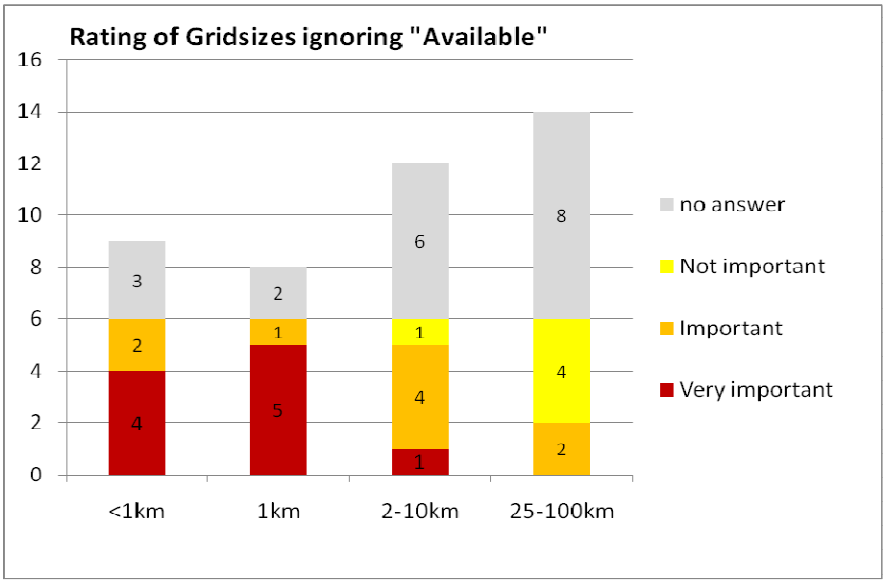


Figure 18: Rating of grid sizes ignoring the option "Available"

9.2.4 Importance of various grid characteristics

National grid systems adopted to suit national Geo-Layers can be constructed with squared grid cells. The INSPIRE – grid was constructed to suit continental Europe and therefore makes transnational datasets possible. However, it may be distorted when it is re-projected to suit national geo-layers. A harmonised grid system for the whole world (e.g. – UTM) cannot have full coverage, but is defined in various zones. Depending on the national projection system chosen, it may again be distorted when re-projected to suit national geo-layers.

The respondents were asked to rate the importance of the following characteristics associated with various grid systems.

- Square grids
- Single grid system for full coverage
- Suits national Geo-Layers
- Suits continental Europe
- Europe-wide harmonisation
- World-wide harmonisation

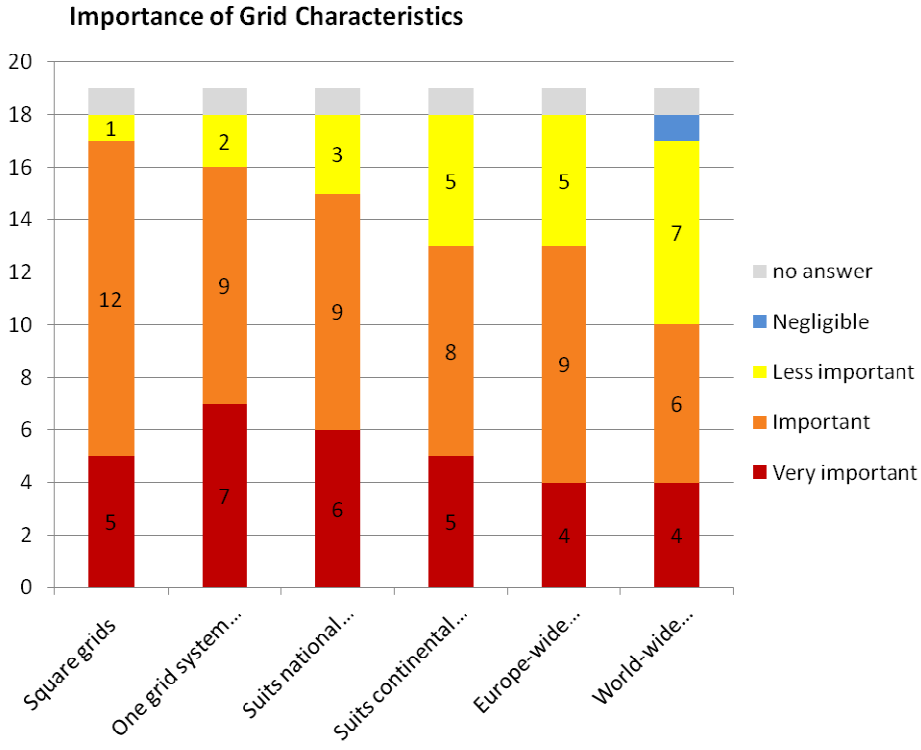


Figure 19: Rating of the importance of various grid characteristics.

Apart from the one respondent who ticked 'very important' for all characteristics, and the three respondents who considered every characteristic as 'important', the answers from the other respondents revealed considerable variations in their ranking of what is most important. This is best shown by the table itself (see Table 7). While five respondents consider square grids as 'very important', only three of them ticked 'one grid system for full coverage' as 'very important'. Of the six respondents who wanted the grids to suit national geo-layers, four are also in favour of a full coverage grid, while for the other two respondents the most important fact is that the grid should suit national geo-layers.

All respondents are in favour of having a grid dataset available which covers the whole of Europe. 53% of respondents (9) consider EUROSTAT as the preferred data provider for a full coverage grid data set of Europe and 41% (7) prefer EFGS.

Table 7: Answers to various grid characteristics

Importance of grid characteristics					
Square grids	**One grid system for full coverage**	**Suits national Geo-Layers**	**Suits continental Europe**	**Europe-wide harmonisation**	**World-wide harmonisation**
Very important	Very important	Very important	Very important	Very important	Very important
Very important	Very important	Important	Important	Very important	Important
Very important	Very important	Important	Important	Less important	Less important
Very important	Important	Less important	Less important	Important	Important
Very important	Important	Less important	Less important	Less important	Very important
Important	Very important	Very important	Important	Important	Important
Important	Very important	Very important	Less important	Important	Very important
Important	Very important	Important	Important	Important	Less important
Important	Important	Very important	Less important	Less important	Less important
Important	Important	Very important	Less important	Less important	Less important
Important	Important	Important	Important	Important	Important
Important	Important	Important	Important	Important	Important
Important	Important	Important	Important	Important	Important
Important	Important	Important	Important	Important	Less important
Important	Important	Less important	Very important	Very important	Less important
Important	Less important	Important	Very important	Important	Negligible
Important	Less important	Important	Very important	Less important	Less important
Less important	Very important	Very important	Very important	Very important	Very important

9.2.5 Confidentiality – Disclosure control

For reasons of data confidentiality in disseminating grid based statistics, there must be restrictions on the disclosure of grid data. Various methods are in use, all of which have disadvantages – either loss of data or loss of spatial accuracy. The most frequent method used

in the respondents' institutions to deal with this issue is deletion (39%), where undisclosed values are replaced by another number or character. Although this method results in a loss of data, the grid size remains the same throughout the dataset.

The method of aggregating grids is used by 28% of respondents. Grid cells with confidential data are merged with other grid cells until the confidentiality threshold is reached. The layer of the data contains grids of different sizes filling up the whole territory. Two respondents chose 'perturbation', a method where undisclosed values are changed slightly by adding or subtracting `noise', one chose "no restriction" and the three remaining respondents chose "Other".

Documented disclosure rules for grid based statistics for the various countries can be found by following these links to national documents:

Denmark	http://www.dst.dk/kvadratnet
Estonia	https://www.riigiteataja.ee/akt/13332259
Austria	http://www.statistik.at/web_en/classifications/regional_breakdown/statistical_grids_etrs_laea/index.html
Finland	http://tilastokeskus.fi/meta/tietosuoja/index_en.html
Norway	Presentation held at EFGS 2010. www.efgs.info
Ireland	1993 Irish Statistics Act; No formulated disclosure rules yet, but they will involve combining a grid with an adjacent grid depending on the population and analysis.

There are basically three main aspects for deciding on what is to be considered confidential and what is not. The thresholds applied when dealing with confidentiality depend either on the variable itself, the number of characteristics into which a variable is split or the size of the grid cell. As combinations of these methods exist, multiple answers were possible.

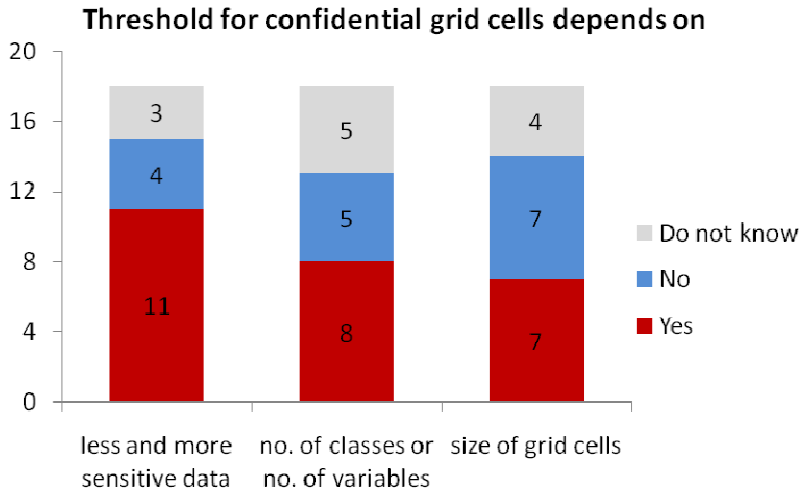


Figure 20: Dependency of threshold for confidential grid cells.

Eleven respondents differentiate between less sensitive and more sensitive data (e.g. information about inhabitants is considered more sensitive than information about buildings), 8

respondents vary between the number of classes (e.g. 5 age classes or 20 age classes) or the number of variables, and 7 vary the size of the grid cells to guarantee confidentiality.

9.2.6 Availability of national grid datasets

Eight institutions prepare grid based statistics in response to customer-specific requests. The following charts show which variables are already available or will be available after the Census/in the near future on the basis of grids, the smallest grid size for which the variable is or will be available, whether disclosure methods are being or will be applied for this variable and giving the most recent year and frequency of update for each variable.

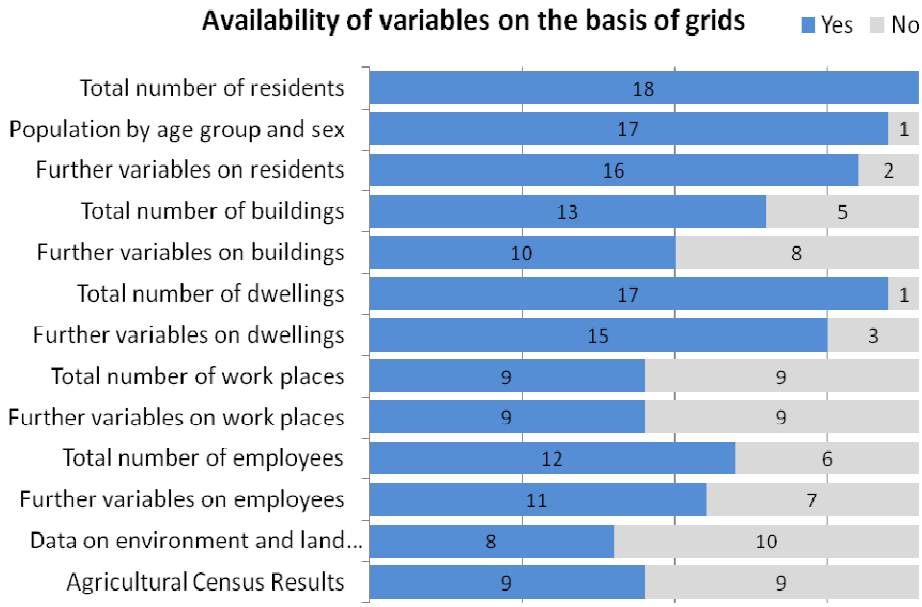


Figure 21: Availability of various variables on the basis of grids.

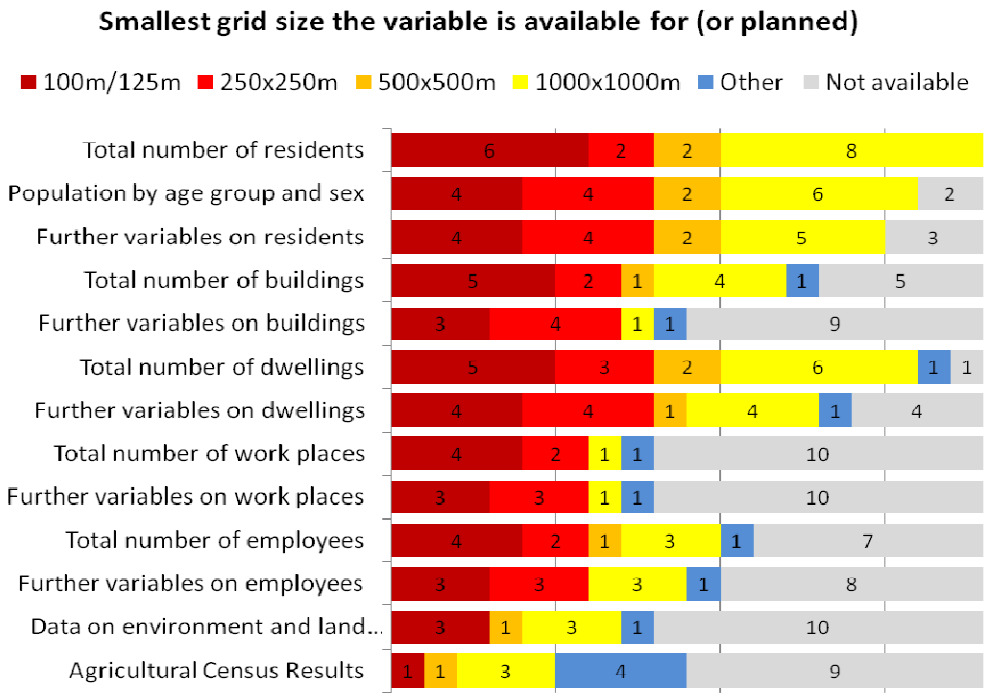


Figure 22: The smallest grid size various variables are available for (or planned).

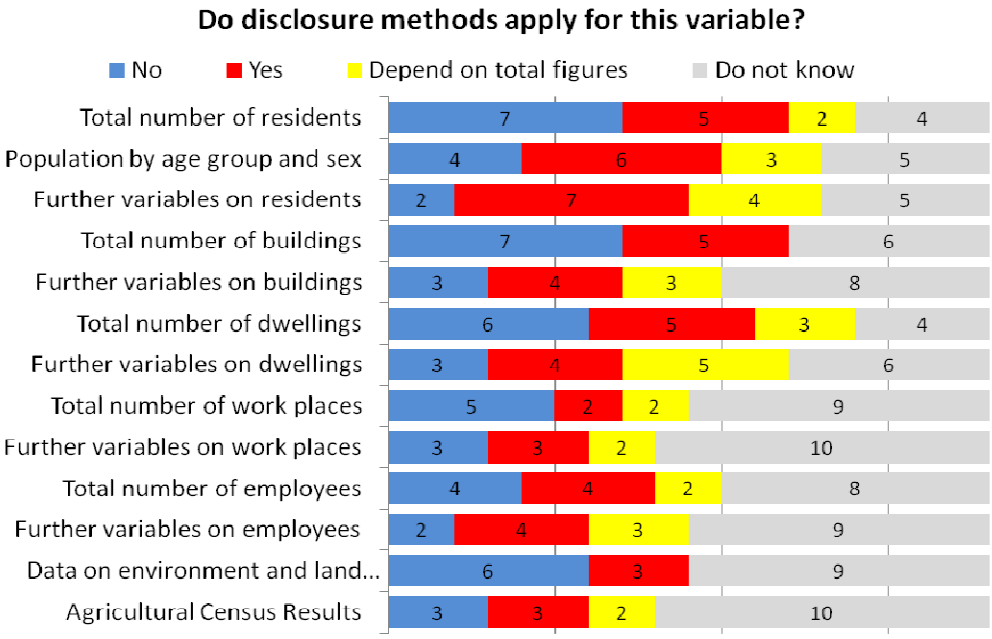


Figure 23: Various variables and disclosure control.

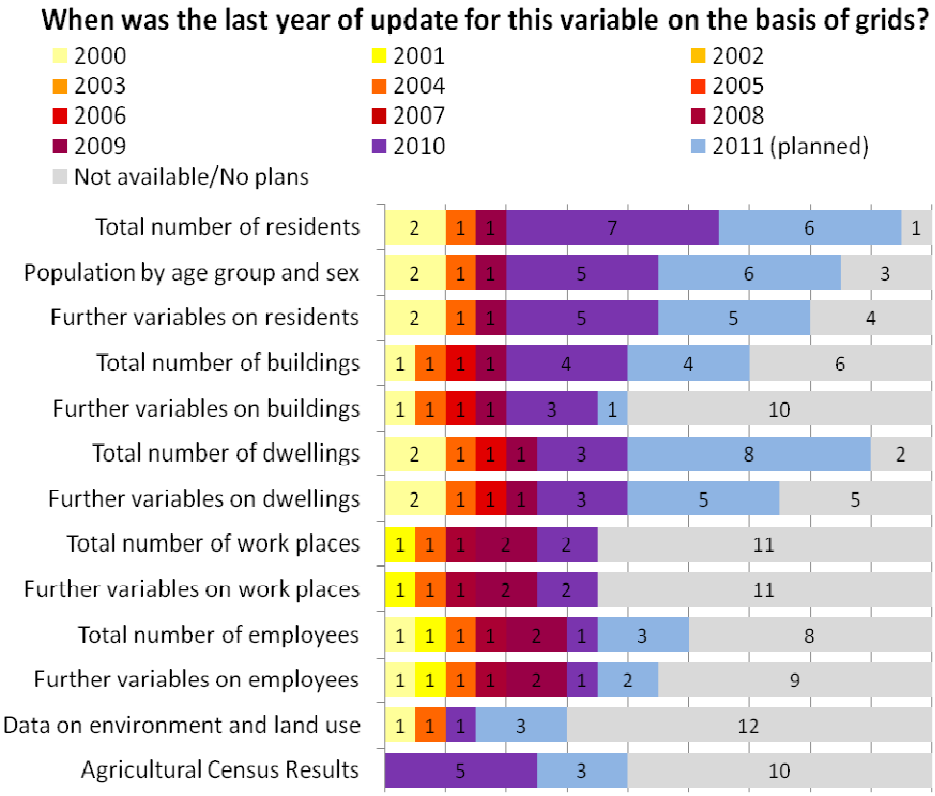


Figure 24: Last year of update for various variables.

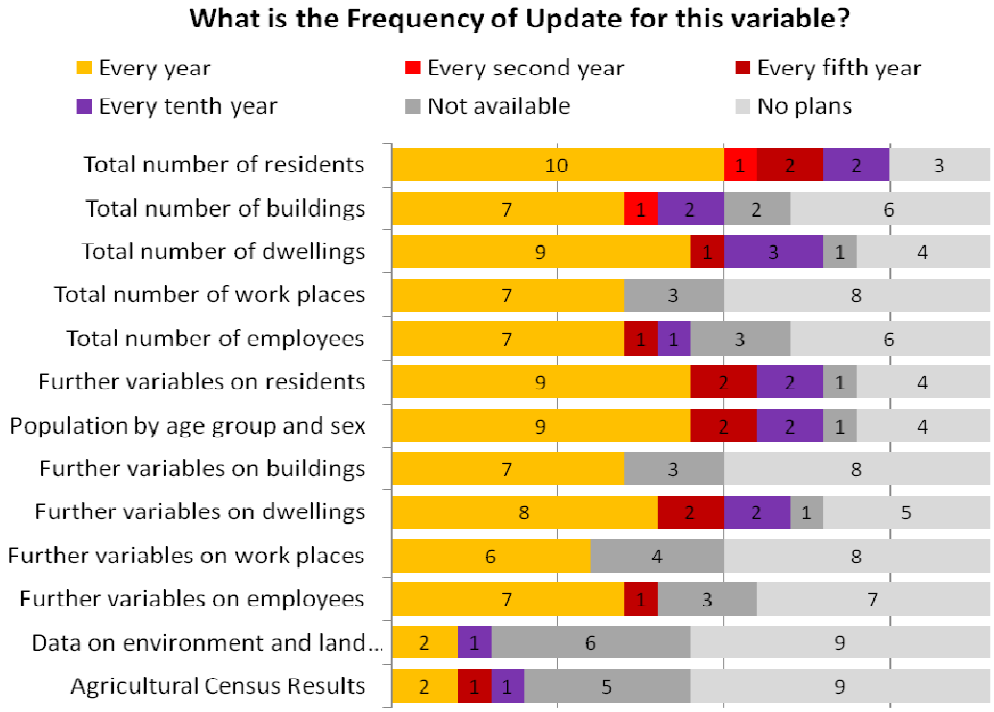


Figure 25: Frequency of update for various variables.

The graphics above show that all respondents have or will have the population figures available on grids (1km or smaller) after the Census. In seven of them the total population

number is available without confidentiality restriction. Most of them also have population figures by age group and sex (89%) and further variables on residents (83%) available on a yearly basis (50%). The total numbers of dwellings (89%), buildings (72%), employees (67%) and work places (50%) are or will also be available in most institutions, but not necessarily in the 1km grid.

9.3 ANNEX III : Data-dissemination survey, results

In a short survey among data providers, the current delivery formats of grid based statistics were investigated, as well as the plans for other types of publication after the 2011Census. A total of 19 respondents coming from 17 countries responded to the data provider survey.

9.3.1 Data delivering file formats

Currently most data files containing grid statistics are delivered in a GIS format, containing geometry (61%); other providers deliver their data mostly in generic text format (22%). Only a minority deliver the data in other formats.

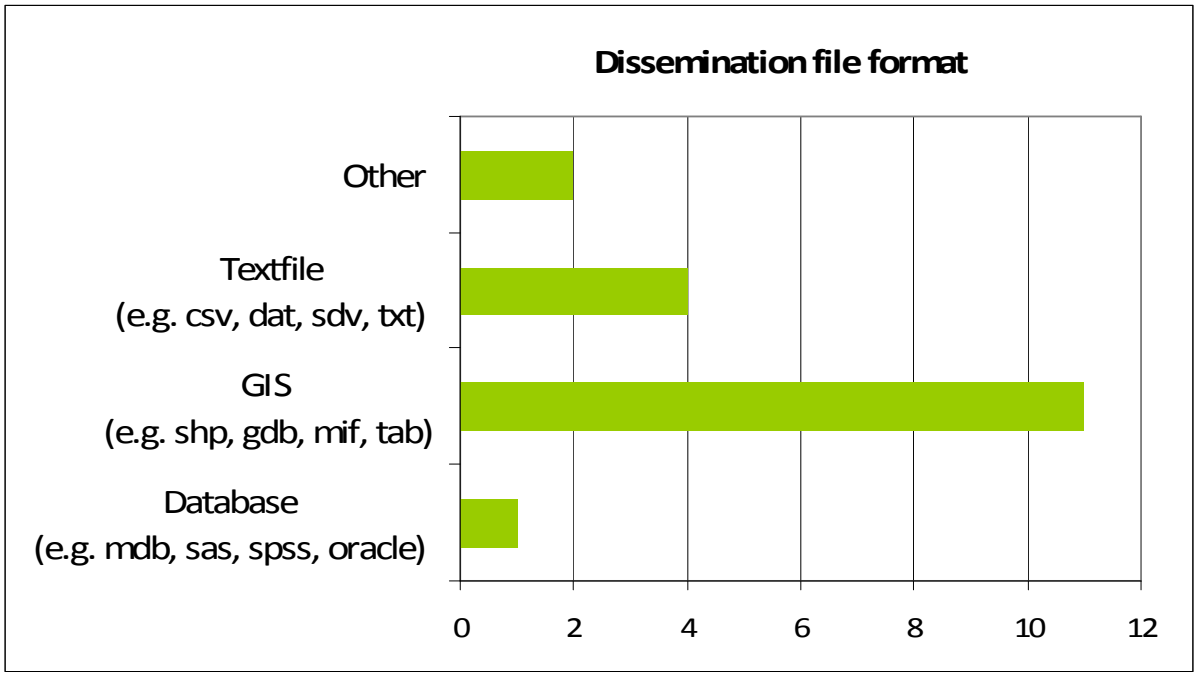


Figure 26: Dissemination file format.

9.3.2 Dissemination of grid statistics after the 2011 Census

Almost all respondents will disseminate grid statistics either by viewing and downloading services or publishing the metadata only. However, it is not clear from this questionnaire how this will be promoted. Although viewing services are mandatory under INSPIRE regulations, only 26% of the 19 respondents are planning to implement WMS services for grid statistics. Statistics by grids will remain available as file downloads (42%), by download service (direct access) for two respondents and by WFS for one respondent. Two respondents have no plans for disseminating grid statistics and two respondents will provide metadata only (multiple answers were possible).

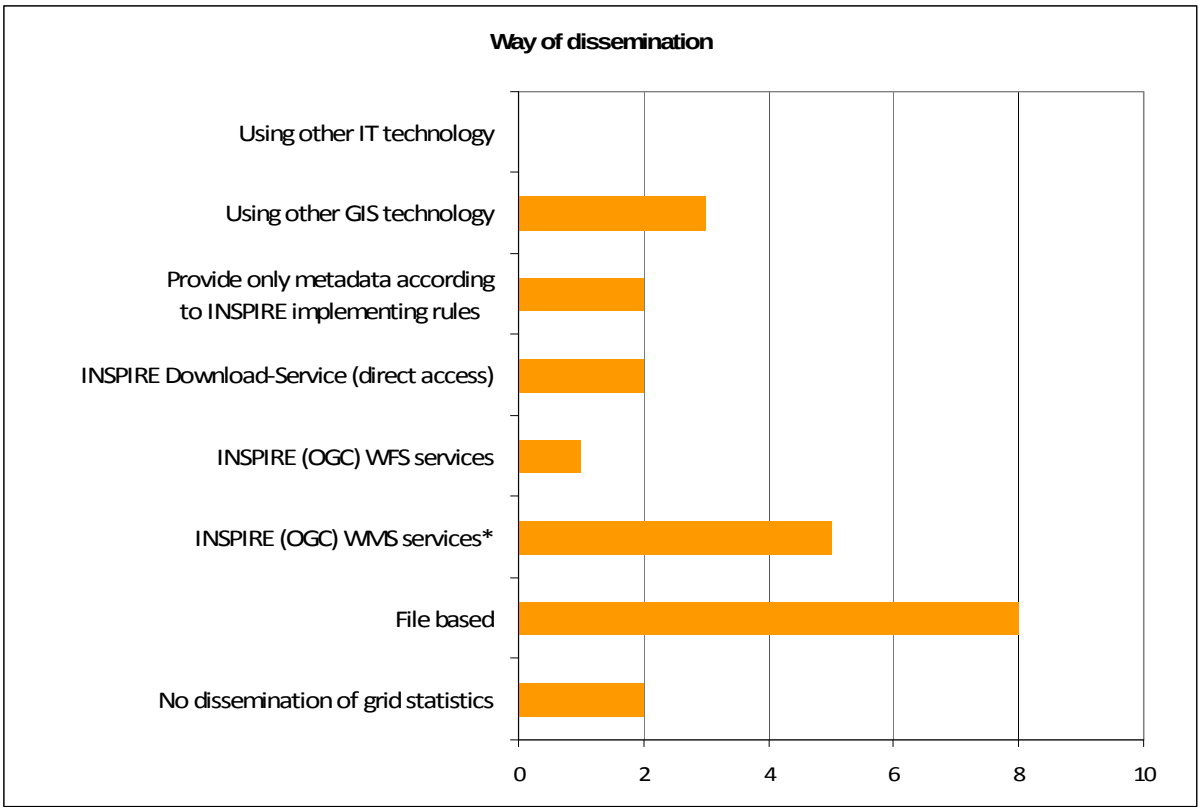


Figure 27: Different ways of dissemination.

9.3.3 Totals for dissemination and access

About two thirds of respondents will make the grid based statistics available after the Census 2011. Three quarters of them (9 out of 12) will disseminate their statistics via viewing services, while 8 out of 12 will make data available for downloading.

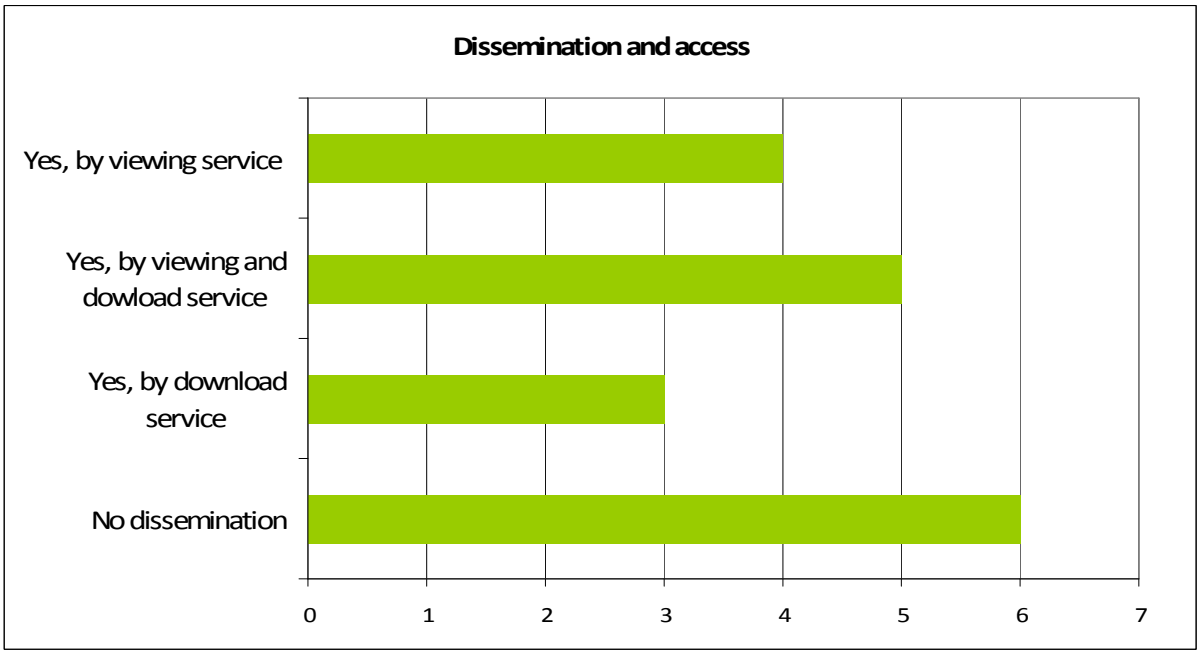


Figure 28: Providing users access to data.

9.3.4 Conditions of dissemination

Conditions for disseminating data, either by viewing services or by downloading services differ between data providers. Six different conditions are distinguished (multiple answers were possible).

- Public domain (The data is available, freely and without restrictions.)
- Reference to NSI required (You give users the right to copy, manipulate, distribute and disseminate the data. In exchange the name of your statistical office has to be mentioned.)
- Only after permission. (Data may only be used when you have been given permission by your organisation.)
- Time limit. (Data is available for a certain period)
- Payment has to be made (Use of data is subject to payment.)
- Data may only be used for non commercial purposes. (Data may only be used within non-commercial applications.)
- Other conditions (If your organisation applies any other conditions)

The conditions for using viewing services have to meet lower thresholds than for downloading services. For almost two thirds of the respondents viewing services will be made freely available without conditions (public domain) or with reference to NSI only, where these more or less free conditions will be met by 8 out of 19 data providers for the downloading of data.

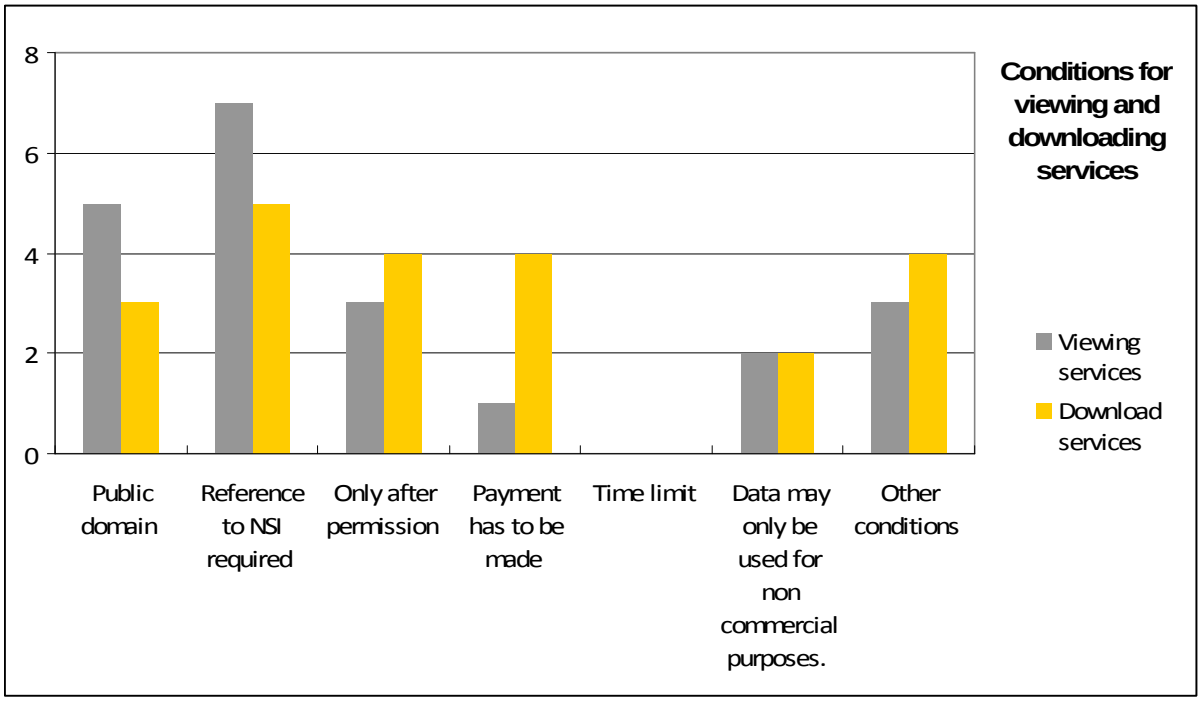


Figure 29: Conditions for viewing and downloading services.

Restricted use by payment is applicable to data downloads in 4 out of 19 cases. The restrictions on the free distribution of data either by payment or by use for non commercial purposes only applies to one third of respondents and is one of the issues to be addressed when data are made available throughout Europe.

9.4 Annex IV: A study of conversions from national spatial data to harmonised European grid data

9.4.1 Point of departure

The aim of the study is to create methods of producing harmonised population distribution data by 1 km square grids according to INSPIRE specifications. The projected co-ordinate system of the target data is ETRS89-LAEA. The specifications for the Geographical Grid Systems have been drafted by the INSPIRE TWG[20] and the actual grid net has been constructed for the whole of Europe as part of the GEOSTAT project[21]. In this study, there is the presumption that national grid data can be constructed or is already available[22]. The aim is to find the best available way to convert national population grid data into harmonised European population grid data.

9.4.2 Differences in locations of grid cells in different projections

Differences in locations between grid nets

The location of a grid cell produced in one projection (or co-ordinate system) is not equivalent to that of a grid cell produced in another system. A grid cell produced using the national ETRS89-TM35FIN co-ordinate system is divided among several ETRS89-LAEA grid cells (Figure 30). The grid net consists of vertical and horizontal lines according to even co-ordinates. The locations of the even co-ordinate lines differ in relation to each other when the projection (or the co-ordinate system) varies. Therefore, the locations of the grids are also different.

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

[21] http://efgs.info/data/GEOSTAT-1km-Grid.zip/view

[22] The guidelines on how to produce population grid data by aggregation, disaggregation or by a hybrid method will be published by the GEOSTAT project later on this year.

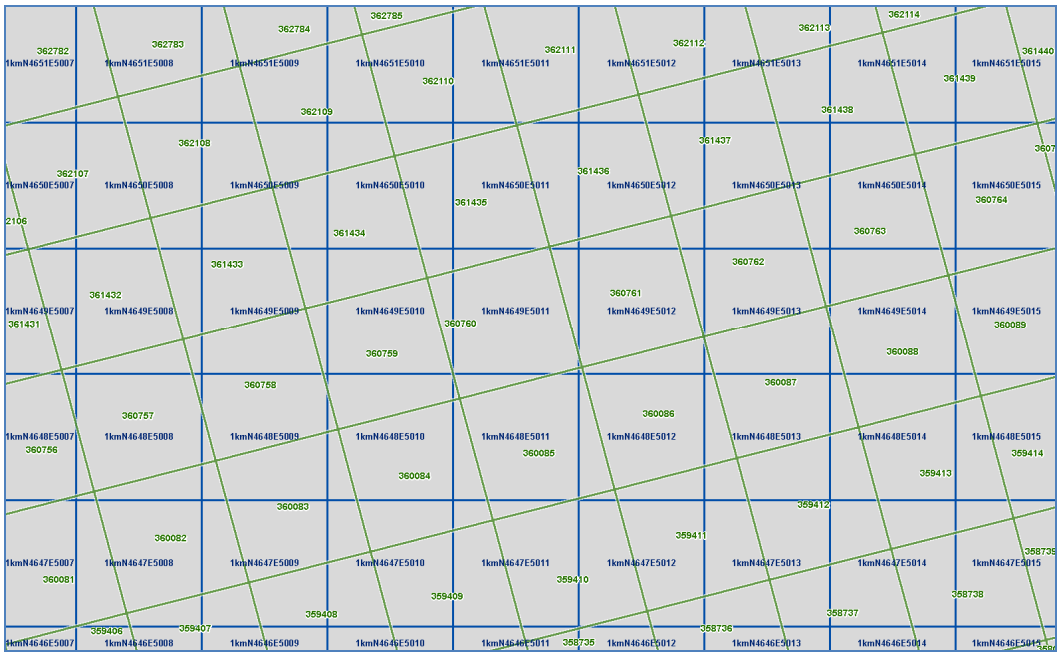


Figure 30: The differences between the ETRS89-LAEA (blue) and ETRS89-TM35FIN (green) grid nets.

Problems with grid net conversions

Like any other spatial polygon, line or point, the grid net itself is convertible. However, quality changes may occur. The grids consist of regular squares, but after a conversion of the projection or co-ordinate system, the shape of the grid cells may become distorted. The target here is not achieved by converting, either. The grid cells in one system do not directly match the grid cells of another system. When overlaying the grid net of one projection on the grid net of another projection there are several intersections between them[23]. It is possible to estimate new/converted grid values using a proportion of the original value to each segment by surface area. However, it was felt that this typr of approach did not fall within the scope of this study.

9.4.3 The methods

In order to produce grid data in a predefined projection (which differs from the projection used at the national level) we describe two methods, one by using building points and another by using ready-made national grid datasets. The test data comprise all inhabited buildings (Method 1) or grid cells (Method 2) in Finland. The tests were made using population data from 31 December 2005 (2006) and 31 December 2010 (2011). The results presented here are from 2005/2006. However, there were no significant differences in the results from year to year.

Method 1: Aggregation of grid data by using converted building points

The first tested method is based on the use of building points (or any other accurate geo-referenced point-based data).

[23] The degree of "matching" is dependent on the relative location of area. E.g., Lambert Azimuthal Equal Area projection in Central Europe is likely to be similar to the nationally used projection in North-Eastern Europe (but still different).

The source data, i.e. building points in the national ETRS89-TM35FIN, were converted into ETRS89-LAEA (Figure 31 and Figure 32). After the conversion, the building point data were joined with the LAEA grid net.

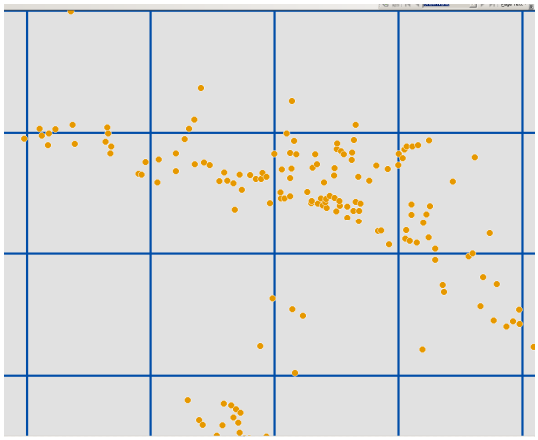


Figure 31 INPUT building points in national ETRS89-TM35FIN projection. ETRS89-LAEA grid net is behind.

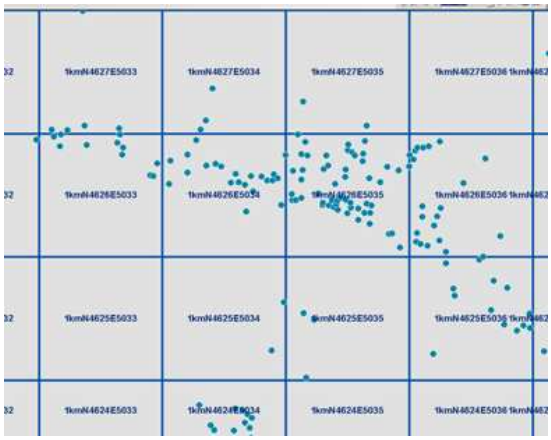


Figure 32 INPUT building points after conversion into ETRS89-LAEA. The location and pattern of the spatial distribution stays unchanged. The converted building points are joined with the ETRS89-LAEA grid net with spatial join.

By spatial join, every building point is given a grid ID, and the aggregation phase can be performed (Figure 33). This method keeps the original data as they are and no quality errors occur. However, the weakness of this method is that it requires the use of primary micro data. Therefore, the method produces double datasets and the national production process will have an equivalent duplicate process for a harmonised European dataset with more or less the same phases.

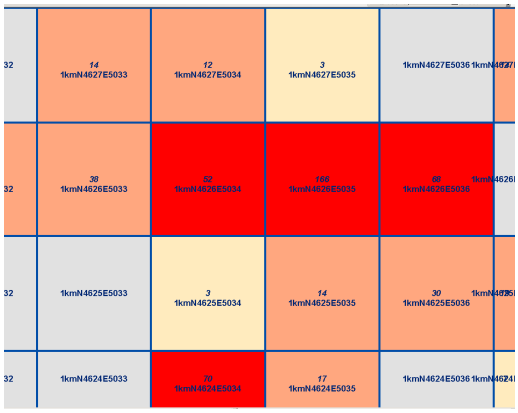


Figure 33: After aggregation, the ETRS89-LAEA grid net contains population sum by grids.

Method 2: Recasting the grids - conversion of grid data starting from ready-made national grid datasets

In the second test, the conversion of grid data into ETRS89-LAEA is performed using national ready-made grid datasets (Figure 34).

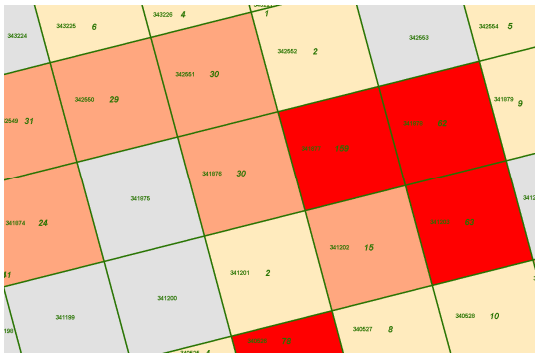


Figure 34: The national grid data are produced by the best national ways.

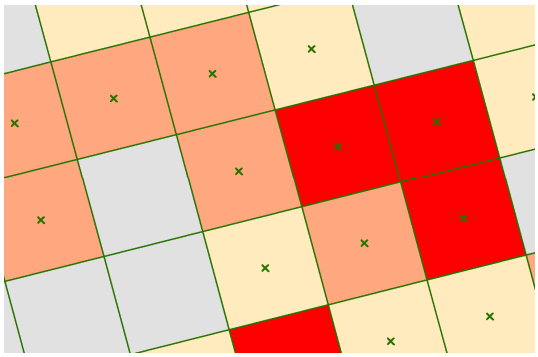


Figure 35: Polygon to point operation creates the central points of the national grid cells. The points represent the grids and all the attribute data a grid cell contains are now the attribute data for the point. The national grid centre points are now converted into an LAEA projection.

In this study, the national ETRS89-TM35FIN grid dataset was converted into ETRS89-LAEA. Given that a direct conversion by reprojection of square grids is not useful due to the distortion effect, the grids were recast instead. Recasting of grid cells involves a number or processing steps. First, the grids were transformed into points. The conversion of the projection is made for the points, which here represent the grids (Figure 35). This test was made by using the middle points of the grids, since it is likely that a middle point best represents the g-ö.rid cell in question[24]. After creating and converting the points to ETRS89-LAEA, the next phase is to overlay the grid points on the LAEA grid net. By means of a spatial join operation, every grid point will have an LAEA grid ID (Figure 36). After the joining, the aggregation is performed (Figure 37). In these tests, the aggregation was made using the Summary Statistics tool in ArcGIS.

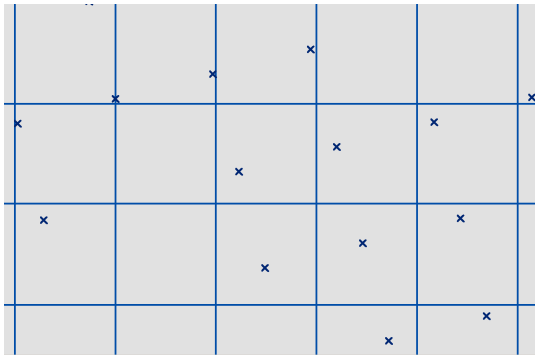


Figure 36: The converted grid points are joined with the LAEA grid net.

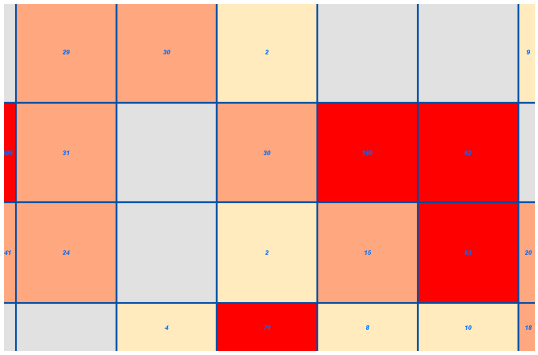


Figure 37: Population by LAEA grid cells.

Effects of the grid cell size on the quality of the recast grid data

[24] When using the GIS software, the polygon to point method usually creates points in the middle of the grid cells. However, if it is felt preferable to locate the points in the lower left-hand corner of the grid cells, they can easily be created by reducing half of the length and height of the grid sides from the original co-ordinates. New points are created from the calculated co-ordinate pairs.

The manner in which the grid cell size of the original national grid data influences the derived one square kilometre LAEA data was also tested. The tests were made using one square kilometre, 250 x 250 metre and 125 x 125 metre grids. The first two sizes are used in the Finnish Grid Database product and in this case represent national ready-made datasets whose data are updated annually. The data by 125 metre grids represent here the finest resolution data available[25]. The data produced by Method 1 (by converting primary data, building points) were used as the reference data (population by one square kilometre in ETRS89-LAEA).

The smaller the size of the grid cell used to recast the dataset the more similar the figures were to the reference data. According to the basic statistical figures, recast data from the 1 km x 1 km national grid data differ the most from the reference data (Table 8). It is worth noting here that the total number of grid cells and their total population are quite similar between the data converted from building points and the data recast from the 125 m x 125 m grid data. The number of grid cells in the dataset is higher when using the recast 250 m grids and lower when using the recast 1 km grids in comparing the reference data. The difference between the 250 m and 1 km datasets in the measured number of inhabitants is due to the differences between the national grid nets of different sizes.

Table 8: Metadata of test datasets - The datasets are: POP_1KM_LAEA - data from converted building points, POP_1KM_125M - data from 125 m x 125 m national grids, POP_1KM_250M - data from 250 m x 250 m national grids, POP_1KM_1 km - data from 1 km x 1 km national grids.

Dataset	Number of grids	Mean	Sum	Minimum	Maximum
Dataset from converted building points					
POP_1KM_LAEA	102 050	51.0	5 204 192	1	14 053
Datasets from converted grid points (by recasting)					
POP_1KM_**125M**	102 249	50.9	5 204 192	1	14 197
POP_1KM_**250M**	102 759	50.6	5 204 166	1	13 283
POP_1KM_**1KM**	99 049	52.5	5 204 179	1	19 175

To a certain extent, the results from the comparisons of the datasets are as expected. The higher the resolution of the national source data, the more accurate the results obtained. The correlations between the test datasets are illustrated in Table 9. All in all, measured with correlation, all the test datasets seem to give very high results. The relevant information here is also the number of matched grid cells. Huge distortion is seen between the recast 1 km grid data in relation to the reference data, with over 20,000 grid cells not matching the reference data. The distortion only shows in a grid-to-grid comparison. As we saw in Table 8, the total population figures remain nearly the same.

Table 9: Pearson Correlation coefficients and number of matched grids between data produced by using building points in LAEA and derived national datasets.

[25] In the Finnish grid data system, 125 m x 125 m grids are only produced to customers' commissions and subject to a user licence in order to limit the use of such data.

	Dataset from converted building points	
	POP_1KM_LAEA	
	Correlation	N of Matched grid cells
Dataset from converted building points		
POP_1KM_LAEA	1.00000	102 050
Datasets from converted grid points (by recasting)		
POP_1KM_125M	0.99900	99 372
POP_1KM_250M	0.99495	97 216
POP_1KM_1KM	0.90989	81 647

The scatter plot illustrations (Figure 38) visualise the level of difference between the recast test datasets in relation to the reference data. The data produced from the 125 m x 125 m grids are very near to the trend line of positive correlation. The 1 km grid data, in turn, are more dispersed[26].

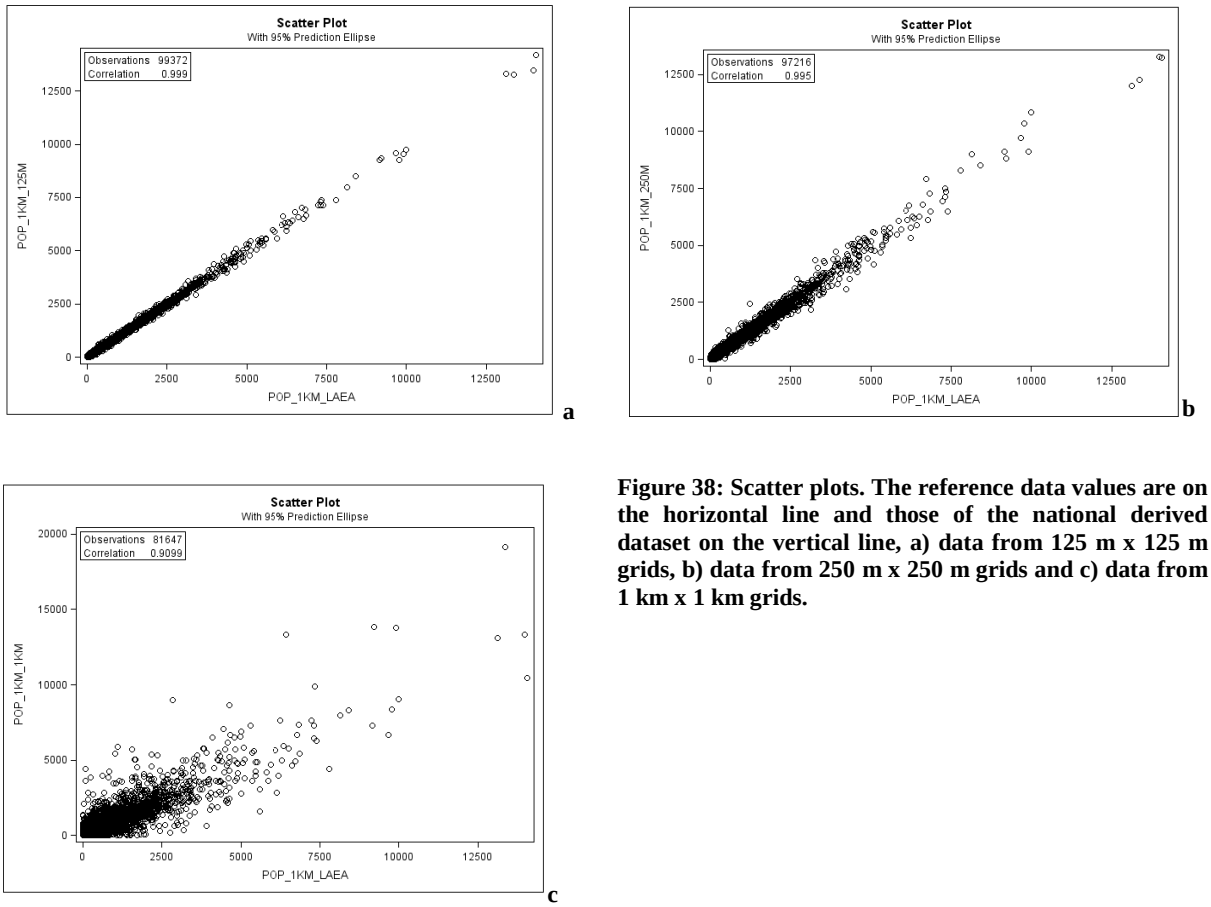


Figure 38: Scatter plots. The reference data values are on the horizontal line and those of the national derived dataset on the vertical line, a) data from 125 m x 125 m grids, b) data from 250 m x 250 m grids and c) data from 1 km x 1 km grids.

The differences between the reference data and the recast datasets were also calculated and the results are presented in Table 10. The differences were analysed using the absolute value of the difference in the number of inhabitants by each grid cell. Comparing the 125 m grids

[26] The scale is different in the x and y axis (see the dash lines, which are drawn to help see the 45 degree identity line).

with the reference data, the proportion of perfectly matched grids where there is no difference at all is 65.7 per cent of all matched grids. The share is 52.2 per cent for 250 m grids and only 24.7 per cent for 1 km grids. If, however, small scale differences are allowed, e.g. the limit of five persons at most, the proportions of matched grids are notably higher when using the high resolution data, namely 91.3 per cent for 125 m data and 85.1 per cent for 250 m data. By contrast, the quality of the 1 km data seems to be quite poor, as the proportion with the difference of 1-5 persons is more than 60 per cent.

Standard deviations of the differences were also calculated. There is a notable increase in the standard deviation figures between the 250 m data and the 1 km data.

Table 10: Differences of derived national datasets compared to data produced by using building points in LAEA. Columns: GRIDS - number of matched grids, Std Dev - Standard deviation of the absolute values of differences, DIF 0 - no difference, DIF 1-5 - absolute values of difference between one to five (inhabitants) , DIF 6-10 - absolute values of difference between six to ten (inhabitants), etc. The differences are expressed also as proportions by classes.

DIFFERENCES (abs.values) between method 1 data (from LAEA buildings) to derived datasets											
	GRIDS	Std Dev	DIF 0	DIF 1-5	DIF 6-10	DIF 11-20	DIF 21-50	DIF 51-100	DIF 101-500	DIF 501-1000	DIF over 1000
125M	99 372	12,7	65 305	25 428	4 429	1 924	1 447	503	335	1	
%			65,7	25,6	4,5	1,9	1,5	0,5	0,3	0,0	
%				91,3							
250M	97 216	28,9	50 742	32 008	7 105	3 156	2 170	1 033	940	56	6
%			52,2	32,9	7,3	3,2	2,2	1,1	1,0	0,1	0,0
%				85,1							
1KM	81 647	135,5	20 194	31 351	11 606	7 839	4 903	1 888	3 000	574	292
%			24,7	38,4	14,2	9,6	6,0	2,3	3,7	0,7	0,4
%				63,1							

9.4.4 Quality and data disclosure

When using a harmonised European-wide grid dataset, the required level of quality depends on the use and the geographical scope of interest. The derived datasets are always slightly fuzzy versions of the original data. The same is also more or less true for the total figures. The inaccuracy is mostly a question of a shift in location. The grid cells located in the neighbourhood have the populations of the surrounded grid cells. When the users' needs are of a general level, the inaccuracy may be irrelevant.

Issues relating to data disclosure have been a subject of concern when producing data in more than just one co-ordinate system or projection. However, it is interesting to note that the method presented here (method 2, use of recasting) does not increase this concern. The original level of spatiality is the character that determines data disclosure. For example, if in the national co-ordinate system (and projection) 250 m x 250 m grids are safe from the point of data protection, no extra information is revealed after the conversion to the grid middle

points. The situation is different when converting the original primary dataset (as building points in the Finnish dataset, method 1). In such a case, a comparison of the produced and the national dataset may expose highly detailed information, which may also contravene the disclosure control rules that have been set. This is also the situation when using the grid middle points if the national disclosure control rules do not allow data with a resolution higher than 1 km x 1 km.

Data quality in country border areas is another issue of concern. By using these methods, the population in the same grid cell can simply be summed up in the datasets of neighbouring countries. Country border areas are not a problem with national, originally aggregated datasets. On the other hand, when this kind of national dataset is next to a disaggregated dataset, double counting of population may occur. However, it may be possible to build a fine tuning method for border areas when developing the disaggregation method.

9.4.5 Summary

Data from converted building points would provide the most qualified harmonised data in spatial terms. However, the double primary datasets and production phases make the method rather heavy. Nor are the primary datasets necessarily available to the grid data producer. Data recasting from the ready-made grid datasets may offer an easier way to reach the target. The smaller the scale of the ready-made source grid data, the higher the standard of the derived grid data can be. National legislation, practices and the nature of the source data determine which method can be used and the quality that can be achieved.

The tests were made using Finnish datasets only. It would also be useful to conduct tests and comparisons with datasets from elsewhere than North-Eastern Europe.

9.5 ANNEX V. Generating population grid data by aggregation - simple tools

Aggregation of data and the computation of final datasets is performed in France using the SAS software alone, which only requires the basic module. The steps are the following :

Step 1 - Build a file of individuals with their geographical coordinates expressed in any national projection system

(in France the “Lambert II étendu” projection system). The actual making of the file is not described here : in France it involves importing coordinates for individuals living in small municipalities and estimating the number of people per dwelling that live in large municipalities. In register-based countries the generation of the individual data may involve importing coordinates or it may require nothing more than just gaining access to the individual file that already has coordinates in it. For the sake of simplicity, the following lines assume that there exists, at the end of step 1, some SAS data file ‘census’ containing a single record for each individual with its coordinates x and y.

Step 2 - Convert the national projection system to the LAEA projection system.

Such a conversion is a simple sequence of mathematical operations. These are well documented on the web site of the French NMA (www.ign.fr). “Algorithms” that can be found there cover a large range of projection systems and therefore may be of some interest to other countries.

Step 3 - Truncate the LAEA coordinates and build the LAEA grid identifier relative to each individual

Step 4 - Aggregate individuals sharing the same grid identifier

Step 3 and 4 can be perfomed in a single step using an SQL statement:

proc sql;

create table grids as select '1kmN'!!put(floor(y/**1000**),z4.)!!'E'!!put(floor(x/**1000**),z4.) as id, floor(x/**1000**)*1000 as x_LAEA, floor(y/**1000**)*1000 as y_LAEA, count(*) as pop from census group by id, x_LAEA, y_LAEA;
quit;

The Output file can be in various formats. The choices made for grid data dissemination on the INSEE’s web site are the DBF and MIF/MID formats.

A first very popular file format is DBF, which is based on an outdated software, but is accepted by a very large number of softwares, including GIS, because of its simplicity. Consisting mostly of human readable text as CSV format, it has the advantage of containing a dictionary of the relevant variables, which reduces the need for documentation.