Geostatistics Portal – an Integrated System for the Dissemination of Geo-Statistical Data

Igor Kuzma[1] | *Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia*

Abstract

A wide range of applicability of spatial statistical data for managing and planning various human activities in the environment or monitoring the trends of different phenomena in space and time requires an adequate response from data providers. The Statistical Office of the Republic of Slovenia (SURS) has a long tradition of processing geo-referenced statistical data that can be point located or aggregated to an optional (administrative) spatial unit and in line with the increasing need for geo-referenced statistical data of high resolution, SURS followed the users' needs by developing various services that are a part of an integrated system for the dissemination of geo-statistical data.

The article discusses the production of geo-statistical data in Slovenia with the focus on the grid data, related confidentiality issues and the system for the dissemination of geo-statistical data, i.e. the Geostatistics portal.

Keywords

Grid data, statistical confidentiality, data visualisation

JEL code

Z, C

INTRODUCTION

Geographical information systems (GIS) have opened a new dimension in understanding of the dissemination of spatial statistical data. To meet the requirements of the growing community of spatial data users, the Statistical Office of the Republic of Slovenia (SURS) adopted new means and formats of spatial data dissemination where grid data proved to be the most challenging. Introducing grids to the standard dissemination process demands a redefinition of the data disclosure rules since grids usually present the phenomena in high spatial resolution. Secondly, the size of the grid data files and the high number of grid cells bearing the data values requires innovative solutions regarding the cartographic presentation and download of the data on the internet.

The entire process of establishing the integrated system for the dissemination of geo-statistical data was primarily focused on creating grid data and the development of the web mapping application that would enable the grid data presentation but parallel to that all data from various statistical domains were

[1] Statistical Office of the Republic of Slovenia, Litostrojska 54, Ljubljana, Slovenia. E-mail: igor.kuzma@gov.si, phone: (+386)2416436.

examined regarding their potential to be included into this system. This cross-examination considered how detailed the particular variable or a set of variables could be presented regarding their reliability and confidentiality. The aim is to present all the data available with the highest spatial resolution possible.

1 GEO-REFERENCING

1.1 Historical background

Register-oriented statistics in Slovenia as expected offered a good foundation for creating geo-statistical data of high resolution. The Register of Spatial Units – initiated by SURS and now managed by the Surveying and Mapping Authority of the Republic of Slovenia – was the first step towards a sound territorial division which enabled geo-referencing (point locating) of statistical data (1971 Population and Housing Census) in Slovenia. These 1971 Census data were used for the establishment of the Central Population Register (CPR) and for the very first time personal identification numbers were assigned to the people residing in Slovenia (Oblak Flander, 2007), which is important for easier later joining of the data from some registers. Although these data could be stored only in tables and not really managed graphically as they can be today by means of GIS, it was decided to permanently preserve the spatial references of the highest possible (or acceptable) positional accuracy.

This far-sighted decision became very relevant when the graphical part of the Register of Spatial Units was completed in 1995. The data stored in tables did have their spatial reference but before that it was very difficult or even impossible to analyse them by means of GIS on the entire national territory. Practically this means that from 1995 on e.g. population data captured in the 1971 Census could be graphically presented for each person on a map at least to the enumeration area (later transformed into spatial districts) of their permanent residence if not down to an address. When SURS started to handle spatial statistical data on grids, mostly the point located data from various registers were considered as applicable, but later also some methods were tested on how to improve the positional accuracy of polygon data while point locating them and aggregating them to grids as described further on. Statistical data from the 1971, 1981, 1991 and 2002 censuses together with the data from the CPR thus offer an important historical picture on how various spatial phenomena changed over the last forty years.

1.2 National hierarchical grid system

SURS has been involved in handling of spatial statistical data on grids since early 1990s with first results of these spatial statistical analyses presented at the end of the decade (Figure 1).

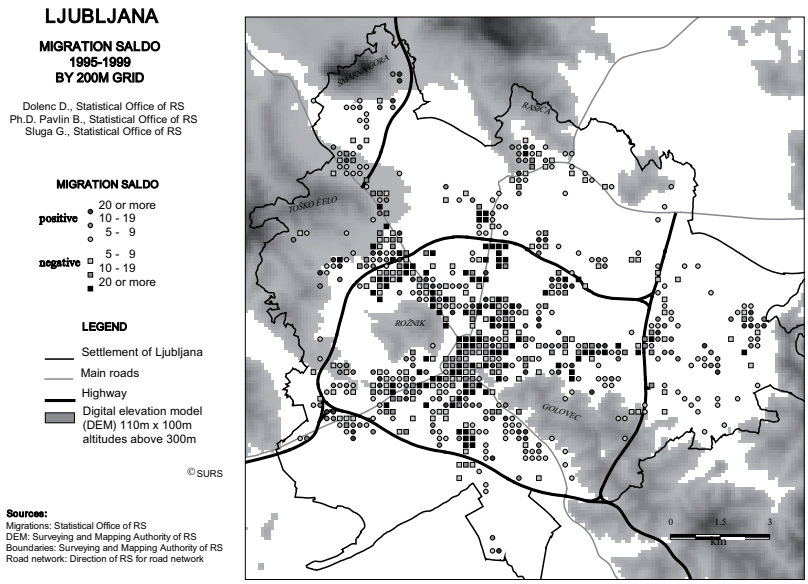
Since then there has been an increase in user demands for statistical data in GIS format, which convinced SURS to further explore the advantages of handling statistical data on grids together with dissemination of such data. Lessons learned from individual case studies and disseminations were obviously to result in the intention to establish a national hierarchical grid system in 2008. Three institutions agreed to cooperate: SURS provided methodological support and the Geodetic Institute of Slovenia together with the Surveying and Mapping Authority of the Republic of Slovenia provided technical support.

The purpose of the joint project was to:

- create square grid vector layers with seven different basic sizes of grid cells,
- define the grid cell nomenclature accordant with the hierarchical structure of grid cells,
- define the origo of the hierarchical grid system,
- define the grid cells both in the previous (D48/GK) and the present (D96/TM) national coordinate systems.

The seven basic grids are 100 m, 200 m, 500 m, 1 000 m, 2 500 m, 5 000 m and 10 000 m grid. The smallest grid cell size 100 m × 100 m was defined considering the user needs for spatial statistical data of high resolution, compliance with other spatial databases in Slovenia and the fact that SURS decided not to present the data for grids smaller than 1ha.

Figure 1 Ljubljana migration balance 1995–1999 on 200m grid



Source: Statistical Office of the Republic of Slovenia

To solve the problem of converting the data from one coordinate system to another it was decided to define square grid vector layers first in D96/TM and then to transform them into D48/GK where grid cells from both coordinate systems share the same cell ID. Transformed grid cells in old coordinate system D48/GK insignificantly lose their square shape but the same cell still covers the same area. Therefore all official spatial statistical data or user's own spatial data which are mostly still in previous D48/GK coordinate system can simply be aggregated to grids in D48/GK and then transformed to D96/TM using the cell IDs. This solution eliminates the problem with overlapping two e.g. same 100m grid datasets defined in different coordinate system (different origo) which might reveal the data values for area smaller than 1 ha.

1.3 Polygons to grids

After establishing the graphical part of the Register of the Spatial Units in 1995 practically all statistical data could be linked to their coordinates and consequently aggregated to optional spatial units including grids. Although the data prior to that could be point located as well by means of various cross-identifiers in reality the greater the time distance the less e.g. persons can be located to coordinates of their permanent addresses. Therefore, some alternative solutions were sought and since the majority of the historical data could still be geo-referenced at least to nowadays spatial districts (originated from enumeration areas) and the fact that the spatial districts are rather small in size their centroids (Figure 2: blue dots) were used as reference coordinates for the aggregation of these data into grids. The centroids of spatial districts are suitable as they are not merely a geometrical centre of the polygon but they mostly coincide with the area of the highest population density in that particular spatial district. These centroids are namely defined by the location of significant objects, e.g. schools. Such centroids can be aggregated to grids as point located data are nowadays, only that the data value for the entire spatial district is geo-referenced to that one coordinate of the centroid of the spatial district. Spatial districts without significant objects obtain their centroids from other significant objects, i.e.:

– centre of gravity of densely built-up area of the spatial district,
– centre of gravity of all buildings in the spatial district when buildings are scattered,
– centre of gravity of the spatial district when there are no buildings in the spatial district.

Any territorial change of the spatial district consequently means a change of its centroid. Despite this, the centroids of spatial districts were additionally examined and corrected where necessary since the population distribution has changed over the past decades significantly in some areas. The correction performed was based on the present state of the centroids of buildings where the information of the construction year of buildings was used to select only buildings which existed and were populated in a particular census period. A common centroid of populated buildings for an individual spatial district is calculated as the average X and Y coordinate of the centroids of populated buildings. Thus a weighted gravity point location of population distribution per spatial district (Figure 2: triangles) is acquired and can be applied to the bottom-up aggregation method (Figure 2).

Figure 2 Aggregation of spatial districts by means of their centroids (circles) or centroids of populated buildings (triangles) into 1 km × 1 km grid

Source: Own construction

In other situations where for instance the number of working places is presented, the centroid of the spatial district can be corrected according to the centre of gravity of all buildings with business activity. Additionally, the position accuracy of 1981 and 1991 census data on population when aggregated to grids can secondly be examined with geo-referenced data from the Central Population Register already available for those periods.

Different from point located data, the polygon data determine the grid cell size according to their average area. Spatial districts in densely populated urban areas cover smaller areas and in rarely populated areas larger areas. Several spatial analyses indicated that the census data can be aggregated to grids with

the cell size 100 m × 100 m or 200 m × 200 m for high, 500 m × 500 m for medium and 1 km × 1 km for low population density areas. Table 1 compares the area of spatial districts to area of 100 m, 200 m, 500 m and 1 km grids together with the number of population. Approximately 47% of the population can be directly aggregated to 100 m, 200 m or 500 m grids thus ensuring high resolution spatial data in densely populated areas.

Table 1 Comparison between spatial districts and grids regarding the area and population distribution

Area of spatial districts in km^2	% of all spatial districts	% of total population	% of national territory	Population / km^2
area ≤ 0.01	11.80	10.85	0.05	20 078
0.01 < area ≤ 0.04	14.28	15.70	0.27	5 770
0.04 < area ≤ 0.25	19.40	20.85	1.90	1 095
0.25 < area ≤ 1	22.15	20.14	11.18	179
area > 1	32.36	32.46	86.60	37

Source: Own construction

The applied methodology suggests that it is highly recommendable to store the census (or other) data together with spatial references of the highest possible positional accuracy if that is legally and technically possible. Positional accuracy can be improved when relevant spatial objects (e.g. buildings) to which most of the data are related are assigned to co-ordinates of their point location and this only has to be done once. A great advantage of geocoded data transformed in this way is that they can be aggregated to an optional grid system regardless of its cartographic projection or coordinate system, but, of course, still considering the appropriate grid cell sizes. Additionally, these data also acquire all advantages of the grid data mentioned above.

2 STATISTICAL CONFIDENTIALITY IN SMALL AREA STATISTICS

Dissemination of geo-referenced statistical data is sensitive since the location of the data is information that could potentially lead to disclosure of the unit of observation, particularly when dealing with data high resolution, e.g. enumeration areas or small grid cells. To assure the confidentiality of the data, SURS defined a set of confidentiality rules that consider sensitivity of variables. Procedures for managing confidentiality are typically produced with coordination of subject-matter specialists, regional statistics specialists and statistical disclosure control experts.

The confidentiality rules differentiate between statistical data geo-referenced to administrative units and grid data. Grid data namely consider the sensitivity of the data and the area of the grid cell size, which should not be smaller than 100 m × 100 m square (1 hectare). Statistical data geo-referenced to administrative units are on the other hand presented regardless of the area of individual administrative unit only considering the sensitivity of the data. Data suppression is used to protect sensitive cells in the attribute table of geo-referenced data sets.

Discussing many solutions already tested and implemented by other NSIs (Strand, Holst Bloch, 2009), it was decided to define the disclosure rules according to the user needs. The analysis of the user requests for geo-statistical data in previous years as expected showed that in case of demographic statistics the users require more spatially and attributively detailed statistical information for densely populated areas where basic figures coming in high resolution were sufficient for remote areas. This fact resulted in determining the non-sensitive statistical data that can be presented without suppression also at the level of settlements and 100 m × 100 m grid cells. These are mainly absolute figures but also sex and age population structure. On the other hand, the sensitive data were determined which are disclosed when the necessary threshold is reached and offer more detailed structure of the attributes. The advanced

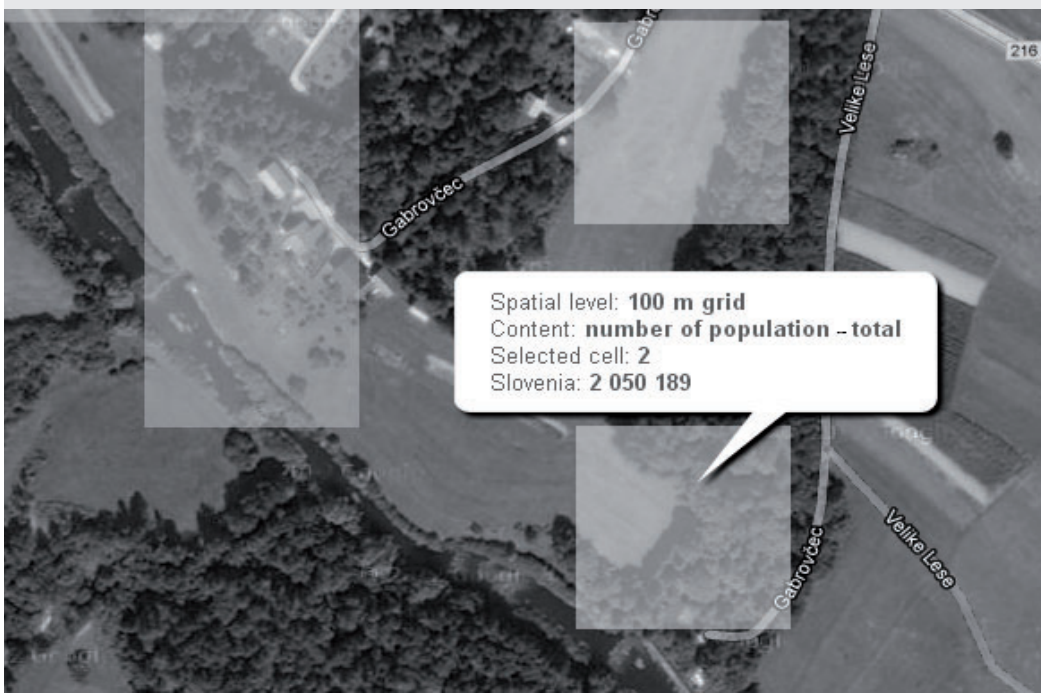
users have as they did in the past still the option to aggregate the spatial units according to their needs, so that the thresholds are met and more information is disclosed for areas where particular phenomena occur in lower density.

The current set of variables available also on 100 m × 100 m grid cells:

- Population:
  - number, sex, age groups,
  - population ≥ 30 → education, activity.
- Households:
  - number,
  - population ≥ 30 → household size.
- Dwellings:
  - number,
  - population ≥ 30 → dwelling area in classes.
- Buildings:
  - number,
  - buildings ≥ 30 → building age in classes.

This method of statistical disclosure was also recognised as optimal since all geo-statistical data are or will be included in various web mapping applications. When combining these data with other geographical information – e.g. satellite images from Google maps – the data may become directly related to individual houses (in case of population data) rather than just the spatial units they are aggregated to (Figure 3).

Figure 3 Web mapping application KASPeR – Potentially problematic presentations of the geostatistics revealing the exact locations



Source: KASPeR, own construction

3 DISSEMINATION SERVICES

The integrated system for the dissemination of geo-statistical data joins various services within the developing Geostatistics portal which will become a common entrance to these types of data and information. The basis of this system is an application for cartographic presentation of the data and download called KASPeR.

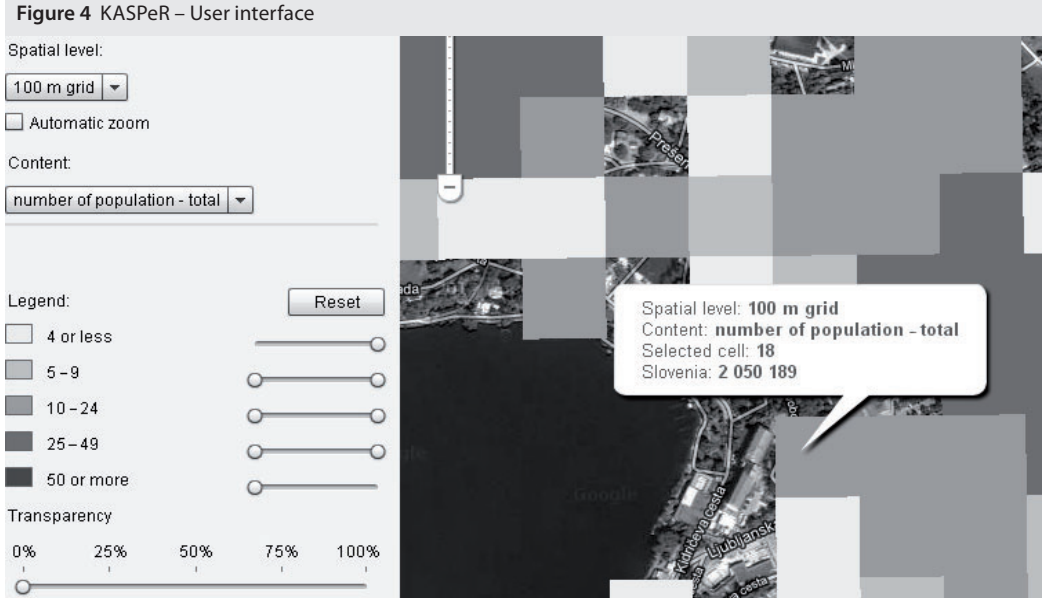
3.1 Visualisation tools

First interactive cartographic presentations of statistical data on the official website were maps created with the PX_MAP, which is a map module within the PC-Axis software family and enables the users to display the selected data in a symbol or a choropleth map. Interactive Statistical Atlas that followed is a Flash based application that allows more interactivity but remains focused on administrative units (NUTS 3, LAU 2). Very popular with the users is a simple application "Place names" presenting the location of the settlements considering the grammatical characteristics of their names. Besides the interactive applications, SURS also offers thematic maps from various statistical domains.

3.2 KASPeR

Web mapping application KASPeR enables visualization of statistical data on different administrative units or grids in combination with maps from the Google Maps tool. The application was designed and developed in cooperation with the Geodetic Institute of Slovenia to explore the possibility of including geo-referenced statistical data of high resolution in a mapping application that would bring the statistics of the 2011 population census to a living environment of an individual. Downscaling from the country level through administrative units to a 100m grid in real time with the help of transparency slider successfully meets that purpose. The application offers a set of demographic variables that can be presented and downloaded as a thematic map. Furthermore, the advanced users are offered a download service that provides free access to geo-referenced statistical data in vector format (*.shp). These data are a valuable input for spatial analyses and data presentation.

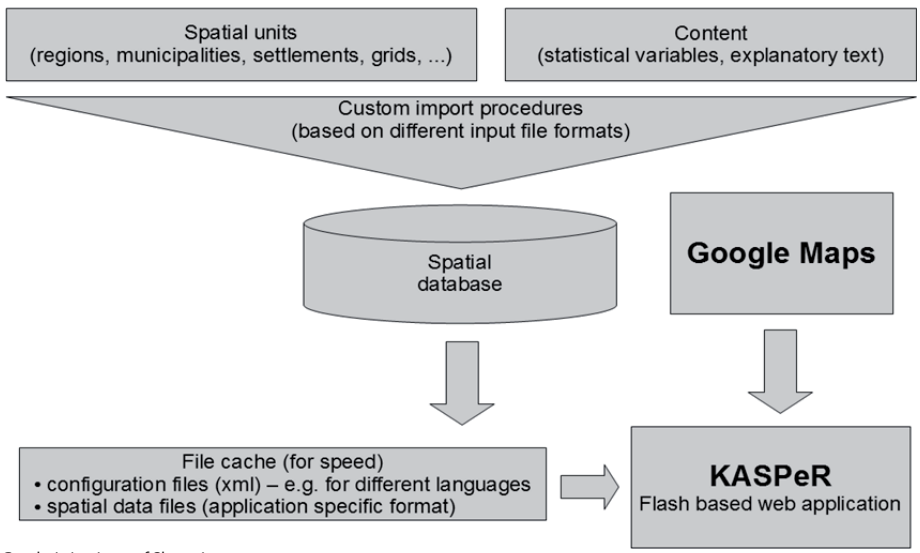
Figure 4 KASPeR – User interface



Source: KASPeR, own construction

The application has a simple data infrastructure with a virtual database supporting the graphical presentations in KASPeR. Spatial units in the form of SHP files are rasterized and organized in a pyramidal hierarchy. Data values are extracted from SHP files and uploaded to a server in compressed data format optimized for quick performance. Administrative units are retrieved from the Register of Spatial Units and assigned the data from the statistical database. Grid data on the other hand are aggregations of point located data with coordinate of the centroid of buildings with an address as a point reference. Since KASPeR is considered more as an experiment, no complex data infrastructure was developed. Consequently, a lot of manual work is involved in the data support, so it soon became obvious that any further improvements of the application should consider automating data input and update.

Figure 5 KASPeR system design



Source: Geodetic institute of Slovenia

3.3 Next steps

Ambitions regarding the future development of the integrated system for the dissemination of geostatistical data at SURS as well as the expectations of the users are great, especially after the successful implementation of KASPeR which is, as mentioned, focused on 2011 census data. The system is being upgraded in the framework of the three tier project performed by SURS and external partners, of which the Geodetic Institute of Slovenia is in charge of the development of a new web mapping application that will substitute the existing KASPeR in 2014.

Tier A covers the development of a new interactive application for viewing and downloading geostatistical data and is co-financed by the Eurostat's grant "Merging statistics and geospatial information in Member States". The new application will cover a significantly larger set of variables from various statistical domains and will include the time dimension introduced by a time slider. A major improvement will be the administrative interface that will enable automated data input directly from the main official statistical database SI-STAT which is built in the PC-Axis environment. Another important new feature of the user interface will be the delineation tools that will enhance the user experience by allowing the users to define their own area of interest either from administrative units, grids or addresses. The application is also expected to graphically complement other SURS' dissemination products that are mainly

presented in textual format. INSPIRE standards regarding the network services will be implemented as well including:

- discovery services,
- view services,
- download services,
- transformation services.

Tier B is being performed by SURS and deals with the upgrade of the GIS database. Improvements will be made regarding the enlargement of the set of variables that will be provided for available time series and geo-spatial datasets defined in two national and one European (ETRS89/LAEA) projections. The set of variables will follow the recently adopted policy that the users should have free access to all relevant statistics also in formats directly applicable to further handling in GIS. Tier C thus demands a thorough study of available statistics provided either by SURS or other producers of official statistics that have a potential and relevancy to be presented as geo-spatial statistical data. Besides the defining of the set of variables, the focus of this Tier is also on defining the confidentiality rules that will determine the level of disclosure as well as the spatial accuracy of the data.

CONCLUSION

Activities described follow the decades of successful work done by SURS and other institutions towards the integration of geography and statistics. Establishing of the Register of Spatial Units and its integration into the statistical process enables various cartographic presentations of statistics and also their application to web mapping visualisation tools. SURS is committed to continue creating new geo-statistical products and improving the system for their dissemination as well as to share the professional experience and integrate national dissemination systems into international infrastructures. Development of a new web mapping application in 2013 that is going to substitute KASPeR will round off the SURS's efforts to establish a user-friendly application that would serve as a viewer and access point for geo-statistics and will be a great upgrade of the existing services joint under SURS's Geostatistics portal.

References

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

Geostatistics portal. <http://www.stat.si/eng/geostatistike.asp>.

Izdelava hierarhičnih mrež Slovenije – zaključno poročilo ob izvedbi projekta. Geodetski inštitut Slovenije, Ljubljana, 2008.

OBLAK FLANDER, A. *Opportunities and Challenges of a Register-Based Census of Population and Housing – the Case in Slovenia* [online]. Seminar on Registers in Statistics – methodology and quality, Helsinki, 2007. [cit. 20.3.2013]. <http://unstats.un.org/unsd/censuskb20/KnowledgebaseArticle10044.aspx>.

STRAND, G. H., HOLST BLOCH, V. V. *Statistical Grids for Norway – Documentation of National Grids for Analysis and Visualisation of Spatial Data in Norway* [online]. pp. 30–32. [cit. 25.3.2013]. <http://www.ssb.no/a/english/publikasjoner/pdf/doc_200909_en/doc_200909_en.pdf>.