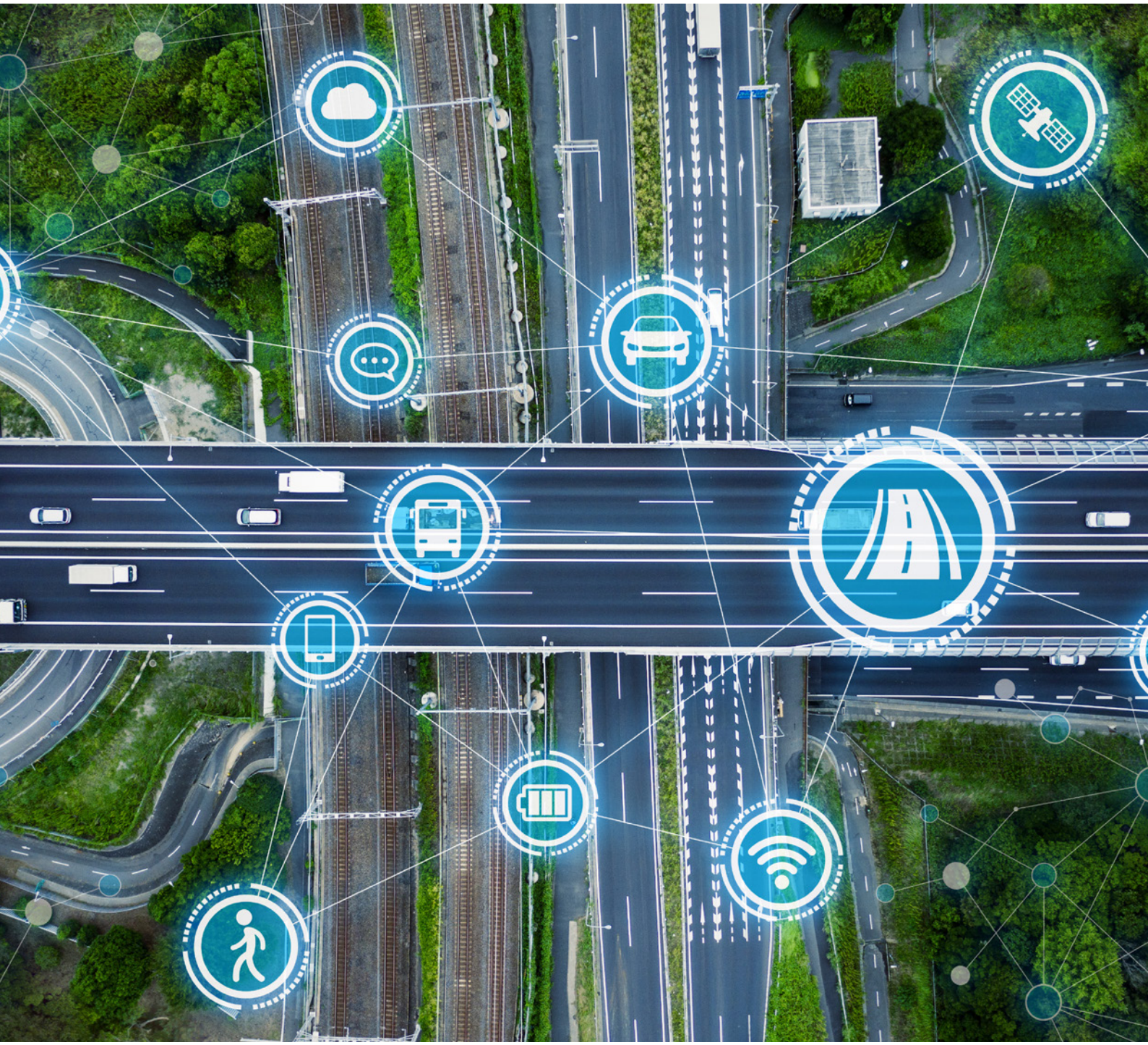
Merging statistics and geospatial information

EXPERIENCES AND OBSERVATIONS FROM NATIONAL STATISTICAL AUTHORITIES, 2012-2015 PROJECTS

2019 edition



STATISTICAL REPORTS

eurostat

Merging statistics and geospatial information

EXPERIENCES AND OBSERVATIONS FROM NATIONAL STATISTICAL AUTHORITIES, 2012-2015 PROJECTS

2019 edition

Printed by Imprimerie Centrale in Luxembourg

Manuscript completed in January 2019.

Luxembourg: Publications Office of the European Union, 2019

Theme: General and regional statistics
Collection: Statistical reports

Print: ISBN 978-92-76-03388-2
ISSN 2529-5233
doi:10.2785/617475
Cat. No: KS-FT-19-004-EN-C

PDF: ISBN 978-92-76-03389-9
ISSN 2529-3222
doi:10.2785/301125
Cat. No: KS-FT-19-004-EN-N

Abstract

Various actions during the course of 2012 — a paper to the European Statistical System Committee (ESSC), a workshop with national statistical institutes (NSIs) and mapping agencies, and a meeting in Prague of Director-Generals of national statistical institutes (DGINS) — resulted in Eurostat deciding to provide a series of grants to statistical authorities to facilitate work on the coordination of statistics and geospatial information.

The association of statistics and geography has the potential to generate information far beyond the simple representation of data on a map. Linking numerical and geo-referenced statistics in spatial analysis may help reveal relationships and phenomena which are difficult to discover by more traditional analyses of statistical databases. *Merging statistics and geospatial information — experiences and observations from national statistical authorities, 2012-2015 projects* presents details of projects enacted with grants provided during the first four years of this initiative, showcasing the broad range of applications that may be developed using geospatial information.

Editor

Hannes I. Reuter

Eurostat, Unit E4 — Regional statistics and geographical information

Contact details

Eurostat
Statistical Office of the European Union
Joseph Bech Building
5, rue Alphonse Weicker
2721 Luxembourg
E-mail: estat-user-support@ec.europa.eu

Production

This publication was produced by Giovanni Albertone, Simon Allen and Andrew Redpath — INFORMA s.à r.l.

For more information please consult

Eurostat's website: https://ec.europa.eu/eurostat/

Acknowledgements

The editor would like to take the opportunity to thank those colleagues in the national statistical authorities who worked on the various grant projects and in particular those who provided final reports for the projects enacted during the period 2012-2015. In addition, the editor would like to thank a colleague in Eurostat who was closely involved in the preparation of this publication, namely Jane Schofield.

Table of contents

Introduction

Location is a key attribute to virtually all official statistics: it provides the structure for collecting, processing, storing, analysing and aggregating data. Moreover, location is a concept that most people are comfortable with, as statistics for a specific place, region or area help people to understand the relevance of particular indicators.

Merging statistics and geospatial information

Traditionally, geospatial information and statistics were attributed to different organisational entities in each country with little cooperation. However, the association of geography and statistics has the potential to generate information far beyond the simple representation of data on a map. Linking geo-referenced and numerical statistics in spatial analysis has the potential to reveal relationships and phenomena which are difficult to discover by analysing statistical databases alone. Furthermore, technological advances and new policy demands have shown that both fields can beneficially be combined. This development has created organisational challenges for the European statistical system (ESS), driving increased levels of cooperation between statistical authorities across the European Union (EU) and other organisations, such as national mapping agencies or providers of big data.

What is a geographical information system (GIS)?

A geographic information system (GIS) is a tool for the management, analysis, presentation and dissemination of geo-referenced data, in other words, data associated to their geographic location. This is evidently the case for topographic information about roads, rivers or administrative boundaries which have been traditionally represented on maps. However, a wide range of additional data sources can also be geo-referenced. Indeed, all statistics inherently have a geographical dimension, be it data covering the whole of an EU Member State, a region, a smaller administrative unit, or indeed an enterprise or a household.

A data revolution has resulted in the ever-increasing availability of statistical and geospatial data. This pattern of development is linked to the growing volume of data that is generated by the internet of things. At the same time as the volume of geospatial data has been increasing exponentially, the ESS has undergone a comprehensive process of reform covering most aspects of its statistical production. These reforms have been driven, in part, by new demands from policymakers to support evidence-based decision-making through better descriptions of societies, economies and the environment within the context of, for example, globalisation, demographic challenges or environmental threats.

Policymaking has increasingly moved across the confines of national borders: examples of current cross-border policies are the Europe 2020 strategy and the sustainable development goals (SDGs). At the same time, European funding for regional and cohesion policy has focused attention on specific territorial characteristics, for example targeting economic, environmental and social problems in cities and/or rural areas. Another change has been the increased level of demand for territorial disaggregation within official statistics: for example, citizens are often most affected by decisions which influence their immediate neighbourhood and this has resulted in governments/local authorities/political opponents increasingly seeking information at a very precise level of detail so they may analyse and illustrate the impact of various programmes and policies. As a result, policymakers and analysts are looking for detailed information across a broad range of spatial dimensions, such as cities and/or rural areas, local administrative units and/or 1 km^2 grid cells.

International background

At a global level, the lead on geospatial information is taken by the United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) who acknowledged the 'critical importance of integrating geospatial information with statistics and socio-economic data'. Their work is based on the development of a global statistical geospatial framework (GSGF) designed to provide an interoperable method for geospatially coding and managing geospatial statistics and information, by connecting statistics that describe socioeconomic and environmental attributes to information that describes our physical man-made and natural environment.

Within Europe, the implementation of a strategy for merging statistics and geospatial information follows global guidelines and has been organised, to a large degree, under the auspices of GEOSTAT projects and annual conferences of the European Forum for Geography and Statistics (EFGS), both of which provide methodological guidance and are funded by Eurostat. The first GEOSTAT project was launched at the beginning of 2010 by Eurostat in cooperation with the European Forum for Geography and Statistics (EFGS), to promote grid-based statistics and more generally to work towards the integration of statistical and geospatial information in a common information infrastructure. Thereafter, there have been two further GEOSTAT projects. GEOSTAT 2 provided a model for a point-based geocoding infrastructure (based on addresses, buildings and/or dwelling registers). GEOSTAT 3 is an on-going initiative, designed to foster better integration of geospatial information and statistics by developing a European version of the global statistical geospatial framework (ESS-SGF) that focuses on providing more qualified descriptions and analyses of society and environment, with work concentrated on sustainable development and the census.

The ESS has responded to these challenges by acknowledging demand for a higher level of geographical detail in its statistics: identifying geospatial information as a valuable data source; recognising its potential for integrating data from multiple sources (administrative, statistical and/or big data); and acknowledging that it represents a considerable opportunity to create more relevant information and respond better to user needs. By geocoding various types of data, statisticians aim to link different data sets through the use of location as a neutral concept, thereby joining data disparate sources together. By doing so, the geocoding of data should help policymakers and analysts to answer the 'where?' in addition to the 'what?' and the 'when?' which have traditionally been the focus of official statistics.

Within national statistical authorities, competences for geospatial information often remains concentrated in specialised teams. They may have considerable experience in using geographical information systems (GIS), with the preparation of statistical maps an established practice. Some EU Member States also use maps to show statistics at a regional level or for smaller-scale territorial typologies. Hence, GIS is by no means a new technology for some members of the ESS. However, given the growing importance of geospatial information, the rapid expansion of data visualisation technologies and the emergence of entirely new groups of users, it would appear all the more important that the expert knowledge embedded in most national statistical authorities is disseminated as widely as possible to colleagues in the same Member State as well as to other Member States. Yet, at the time of writing, the work conducted by a relatively large number of statisticians has yet to be impacted by developments in GIS, with many having few or no specific competences in this area.

The focus of geostatistics is principally on spatial things, in other words, real-world phenomena that have spatial extent or position. A spatial thing has various characteristics (such as its shape, name or boundary). Spatial things can be material (such as monuments, buildings or bridges) or non-material (such as administrative boundaries; boundaries of cadastral parcels of land; routes within a transport network). To code data at such a precise level, an address is often required: this provides a geographic data item that is used upstream of statistical processes, for example, during the collection of information either by post or in the field (by an interviewer). In both cases, an address provides direct access to the required respondent or object (by referring to a precise location with geographical coordinates that eliminates any ambiguity).

What is GISCO?

The geographical information system of the (European) Commission — GISCO — contains essential datasets such as topographic data and political boundaries. It provides corporate data and services for: administrative and statistical areas; hydrography; transport; land cover/land use; population distribution. These are areas of common interest for multiple stakeholders and are regularly used by a range of European Commission services that focus on regional policies, for example, the Directorate-Generals for Regional and Urban Policy, Mobility and Transport, Environment, Energy, Agriculture, Maritime Affairs and Fisheries, or the Joint Research Centre (JRC), as well as the European Environment Agency (EEA).

Data included in GISCO are defined by the GIS user community, organised through the European Commission's inter-service group on geographical information (COGI); they ensure the coherence, consistency and usability of data. The COGI also ensures the consistent and effective use of geographic information across European Commission services, as well as coordinating elements such as data acquisition, its software portfolio, the sharing of information and expertise, as well as the implementation of INSPIRE (see below) within the European Commission. Note that most of the data in GISCO are not available to the general public due to copyright/license limitations.

An effective geo-referenced statistical information infrastructure should be consistent and interoperable with spatial data infrastructures developed following the INSPIRE Directive (see box below). Indeed, more accurate and better exploitable geostatistics may generate added value in a range of areas:

- statistical authorities may reduce the cost of data collection, for example, by using GIS techniques and location-enabled devices to plan surveys better;
- statistical collection could be made more efficient by exploiting geo-referenced data and/or making use of it when there were changes for reporting requirements;
- dissemination could be improved, by aggregating the same data to produce statistics for, among others, grids, functional areas, regions or river basin districts.

What is INSPIRE?

The INSPIRE Directive (2007/2/EC) entered into force in May 2007, establishing an infrastructure for spatial information in Europe (INSPIRE) to support Community environmental policies, and policies or activities which may have an impact on the environment. Its goal is to make geographic information held by public administrations more accessible through a geoportal (http://inspire-geoportal.ec.europa.eu/) that is accessible to everybody. To do so, data and metadata across 34 spatial data themes from regional, national and international sources are harmonised using an agreed set of standards that make it possible to share, combine and aggregate spatial information.

The INSPIRE infrastructure is characterised by the integration of spatial data from multiple sources across socioeconomic themes, for example, covering statistical and administrative units, population distributions, health statistics or information on energy and environmental resources. INSPIRE recommends the use of table joining services to integrate socioeconomic data with geographic data for administrative boundaries and/or statistical units. Simply put, this means linking information about people, businesses and physical objects to a particular place in order to improve the understanding of complex social, economic and environmental issues through data analysis, spatial analysis and thematic mapping.

More information is available at: https://inspire.ec.europa.eu/

Investment in geostatistics also has the potential to improve greatly the synergies between national statistical authorities, mapping agencies and other authorities dealing in geographic and cadastral information: promoting the exchange and integration of statistics; avoiding duplication in data collection; and facilitating European, national and subnational reporting obligations. In order to be as effective and efficient as possible, the joining together of statistics and geospatial information needs to be focussed on harmonisation at the start of the process. In this way, geo-referenced statistics at microdata level can later be fully exploited, allowing them to be used in more flexible ways and for a broader range of analyses. For example, a single set of microdata may be used for providing statistics across a range of different territorial typologies, while the same data may also be used to analyse issues which are not necessarily known at the moment a survey was designed/carried out.

Such changes in statistical production have led to a number of established statistical conventions and rules being re-examined, the most prominent of which is the question of confidentiality. Official statistics have traditionally put considerable efforts into making the identification of statistical entities impossible or at least very difficult; this has mainly been done by eliminating confidential data from published results or aggregating data to such a level that prevents disclosure of the sensitive records. Geospatial information has a number of specific considerations and requires different types of safeguard to protect confidential information; the increased use of registries, disaggregation techniques and methods for modelling small areas are some of the most relevant areas of debate. Otherwise, the use of richer, more specific and detailed information also raises issues around business models and commercially relevant data, for example, those associated with common licencing schemes.

Purpose of this publication

Since 2009, Eurostat and the ESS have stepped up efforts to include detailed location or functional geographic classifications as an important parameter in various environmental, social and economic statistics. This goal was designed to enhance the information capacity of statistical data, mainly for planning, programming and spatial analysis, without increasing the cost of creating the data; the main driver for this initiative was the 2011 census round.

During the course of 2012, three different events were organised by the European Commission in relation to statistics and geospatial information:

- a paper was presented to the European Statistical System Committee (ESSC);
- a workshop was organised between the European Commission, national statistical authorities and mapping agencies;
- a meeting was held in Prague of the Director-Generals of national statistical authorities (DGINS).

Eurostat subsequently sought to promote integrated information systems that combine statistical and geospatial information for policymakers, researchers, spatial planners, as well as a range of other users, sharing this knowledge with the wider community, providing an overview of how geographical information systems have been implemented and identifying issues for further guidance and future developments.

The three events referred to above led Eurostat to launch a call for proposals in 2012 under the heading of *Merging statistics and geospatial information*. This was designed to provide grants for facilitating work on the coordination of statistics and geospatial information. It was intended to cover a wider range of topics, including:

- improving the integration of geographic information and geo-referencing in the statistical production process;
- illustrating how linking geographical and statistical information provides additional value and creates new information;
- designing innovative web applications to show the spatial distribution of statistics.

The call was also intended to increase cooperation between national statistical authorities, in the sense that tools or processes designed or developed by one Member State might be offered to others for reuse and/or inspiration. Furthermore, national statistical authorities were explicitly encouraged to propose projects together with organisations responsible for geospatial information, in particular, national mapping and cadastral authorities (NMCA) to promote greater cooperation and a cross-fertilisation of information.

This publication, *Merging statistics and geospatial information — experiences and observations from national statistical authorities, 2012-2015*, presents details for each of the projects provided a grant during the first four years, showcasing a broad range of applications that may be developed using geospatial information.

For the first exercise in 2012, Eurostat received 11 proposals and selected eight of these for grants, namely those from Greece, Hungary, Malta, the Netherlands, Poland, Slovenia, Slovakia and the United Kingdom. Some of the projects were cross-cutting and ranged from data collection to web dissemination, while others were focused on a specific aspect of the business process in an individual statistical authority. All projects were thought to have enhanced the GIS expertise of NSIs and made substantial progress in giving increased visibility to GIS. However, there were concerns raised as to the potential transferability of projects between NSIs, while no projects were carried out in unison with national mapping and cadastral authorities. In response, Eurostat set-up a collaborative platform to exchange information and promote reusing results from other countries. Furthermore, it was agreed that future calls should promote projects that sought to bring location into the mainstream of statistical production (by developing geocoded data warehouses) and help NSIs prepare for the 2021 census exercise (through the implementation of geocoded data, covering buildings, addresses, citizens, businesses, workplaces and farm holdings, thus creating a point-based framework for statistical microdata).

For the second exercise in 2013, there were seven projects selected by Eurostat for grants, namely those from Bulgaria, Germany, Croatia, Italy, Austria, Slovenia and Finland. One of these continued work that was started under the 2012 grant, whereas three others were centred on extending national capabilities for merging statistics and geospatial information. There were also three more specific projects that were focused on: manipulating information collected from migrant arrivals (collected when they applied for a residence permit), so that their place of birth could be geocoded, allowing a set of 20 maps to be produced for non-EU countries, showing the precise origin of migrant arrivals for the years 2012-2015. Another project allowed a set of commuter statistics to be developed, providing information on the average distance commuters travelled to work or to their studies. The final project was more closely linked to information technology, namely, developing an open source web application for the spatial analysis of statistical data.

For the third exercise in 2014, there were seven projects selected for grants, namely those from Estonia, France, Croatia, Hungary, Poland, Portugal and Norway. These covered a broad range of issues including: linking statistical registers to address systems; producing multi-modal spatial transport data for urban centres; assessing how changes in population and land use may impact on the quality of life; or establishing a point based business register.

For the fourth exercise in 2015, there were eight projects selected for grants, namely those from France (this project was extended until 2018), Croatia, Latvia, the Netherlands, Austria, Poland, Slovenia and Finland. Note that the grant provided to the French national statistical authority concerned the preparation of a methodological handbook. As such, it did not specifically cover a practical application for merging statistics and geospatial information and for this reason has not been included in the main body of this publication. Nevertheless, the handbook produced provides a very valuable tool that may be used to promote and share results, encouraging a greater take-up and application of spatial statistics in statistical production processes. The core of the handbook focuses on describing geocoded data, measuring the importance of spatial effects, describing practical methods for taking into account spatial interactions and providing details on some more advanced issues and latest developments (spatial panel data models, network analysis, spatial econometrics, small area methods). The *Handbook of Spatial Analysis* is available in both French and English.

Glossary

Address is the specific location of a property, usually based on address identifiers such as a road name, house number and/or postal code.

Continuous data describes data where values for the variable of interest may be observed at any point across the territory studied. Data are generated on a continuous basis, but they are measured only at a discrete number of points (for example, the chemical composition of the soil, water or air quality when analysing land use or land cover.

Geocoding is the process of transforming a description of location (such as an address or the name of a place) to a location on the Earth's surface. Geocoding is the process of linking unreferenced location information, often in the form of a text string (an address) to a geocode. The conditions for geocoding include a high quality physical address, property or building identifier, or other location descriptor, in order to assign accurate coordinates and/or a small geographic area to each statistical unit.

Geographical classifications are methods to group geographies according objective criteria, for example classifications based on population density, functional aspects (labour market areas), or geography (mountain areas). Often geographical classifications are based on statistical or administrative geographies to be able to compare statistics between different areas with the same characteristics (for example, urban areas).

Geo-referenced statistical data are data that can be directly presented in space. Geo-referencing, or geospatial referencing, is the process of referencing data against a known geospatial coordinate system, by matching it to known points of reference in the coordinate system, so the data can be processed, queried and analysed with other geographic data.

Geospatial core information is a set of essential geospatial data and services for geocoding other types of information; examples include administrative boundaries, land cover information, addresses, orthophotos/satellite images, transport and hydrographic networks.

Geospatial data are information defined by geometrical boundaries of either administrative or other units that are in geographic information systems (GIS), commonly in the form of polygons.

Grid statistics are spatial statistics geocoded to rectangular grid cells. Each grid cell has the same size and carries a unique code. Ideally the code carries also geocoding information, for example, the lower left corner of the grid cell.

Linking defines a process of connecting structured data sources using a system of unique identifiers. While integration describes the process of combining data from different thematic communities, linking refers to technically connecting data in a machine-to-machine environment.

Location is a general term used to describe a place on the surface of the Earth; location data is often used when referring to geospatial information.

Point data are those whereby the geographic coordinates are associated with an observation. The value associated with the observation is not of interest, rather it is the location, for example, the point where a disease emerged during an epidemic, or how certain tree species are distributed. Spatial analysis of point data is aimed at quantifying the gap between observations, identifying clusters of data that are more aggregated than if they had been randomly distributed across the territory.

Regional statistics are statistics that are geocoded to administrative and functional geographies.

Spatial analysis or **spatial statistics** include any of the formal techniques which study entities using their topological, geometric or geographic properties; the phrase refers to a variety of techniques, many still in their early development, using different analytic approaches applied in a wide variety of fields.

Spatial data are statistics which should meet the perception of users in their area of interest (for example their neighbourhood); as such, these data are more detailed than regional statistics. Spatial statistics are geocoded to small administrative or non-administrative geographies.

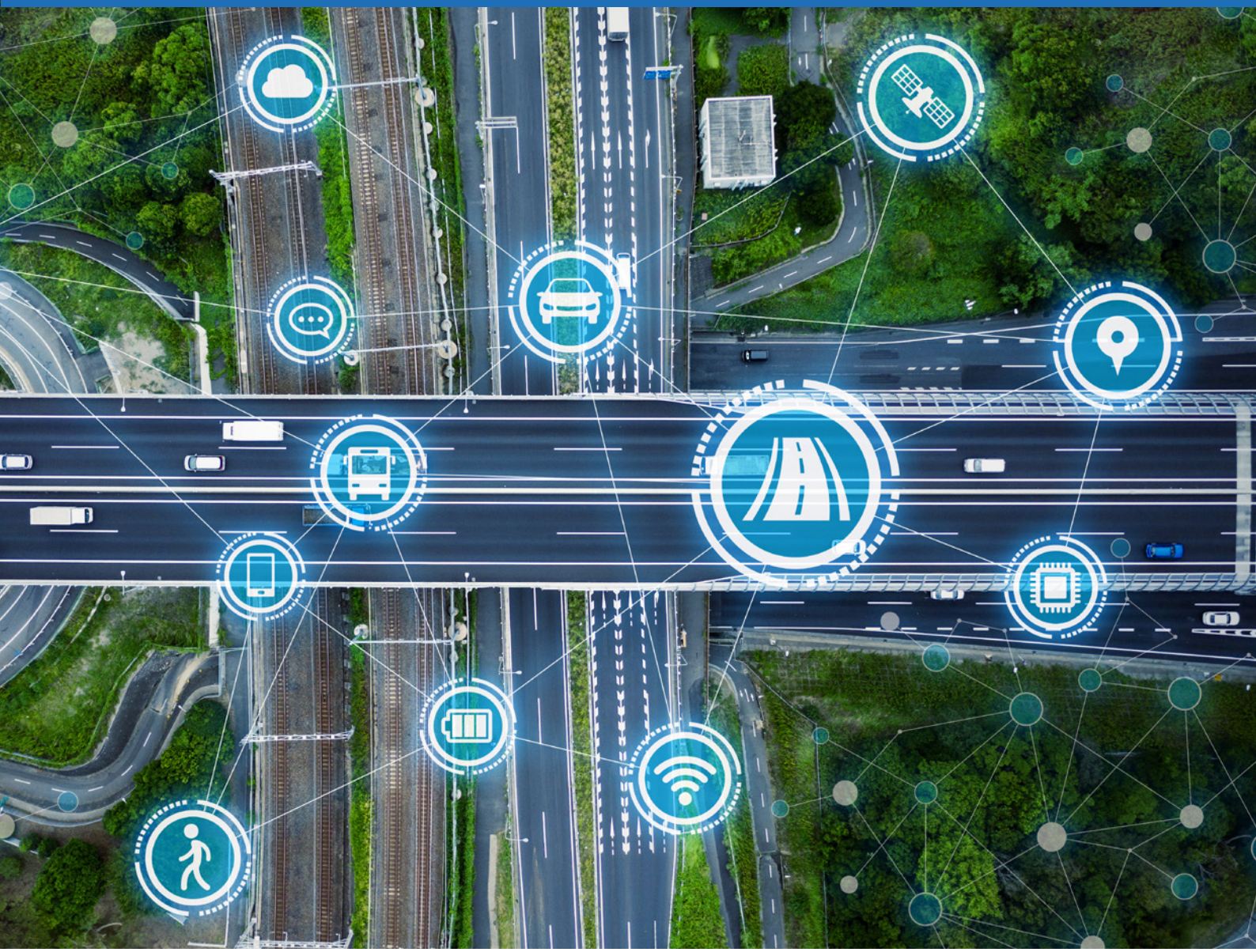
Spatial unit refers to any set of spatial units that cover a whole territory and are divided into basic (house number, spatial district, statistical district, settlement, municipality, administrative unit) or additional spatial units (local, village or urban community, street, electoral unit).

Statistical unit describes one member of a set of entities being studied; this could include persons, households, businesses, buildings or parcels/units of land.

Thematic map is a type of map or chart that is designed to show a particular theme connected with a specific geographic area; these maps can portray physical, social, political, cultural, economic, sociological or agricultural patterns and developments.

A

2012 projects



Greece

Hellenic Statistical Authority

Compilation of a Greek population grid for the year 2001 and web mapping design and implementation; 2012 project; final report 17 June 2014

KEYWORDS: demographics, census, grid, visualisation

PROBLEM

A lack of detailed information for the distribution of the resident population, as census data could only be georeferenced at territorial levels of variable sizes.

OBJECTIVES

The first objective of the project was to develop a 1 km^2 population grid for Greece based on the 2001 census data and to put in place a system/infrastructure that could be used to treat information from the 2011 census exercise.

The second objective was to develop a web mapping application.

METHOD

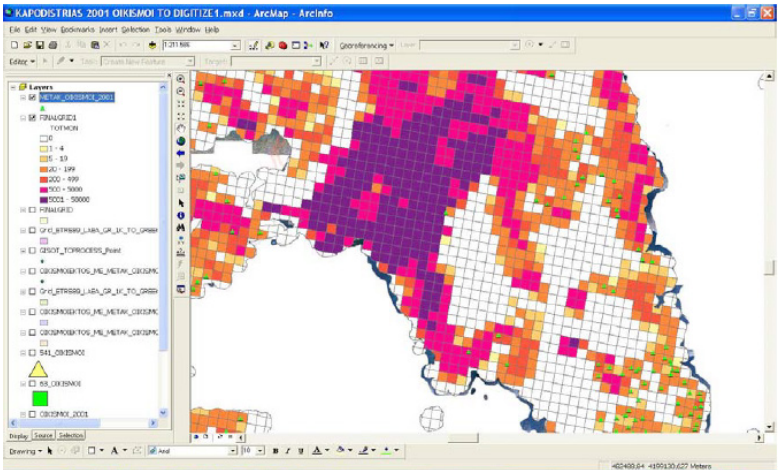
In order to create a population grid based on 2001 data, the project foresaw using a bottom-up approach to create a more realistic depiction of the Greek resident population.

The census information did not include point data (information for buildings) and there was not a registry of buildings/addresses. Furthermore, data at the level of the census blocks were not available for all settlements. As a result, it was decided to focus on the creation of a 1 km^2 population grid, rather than any finer level of detail (for example, 500 m * 500 m or 250 m * 250 m).

The project used several sources of data: the Greek portion of the European population grid, census block polygons for 604 municipalities and a digitized polygon dataset showing the area of 12 092 inhabited settlements. Orthophotographs (aerial photographs) from the national cadastre and mapping agency were used for digitising and for verification purposes, as was Google Earth for verification purposes.

The data processing involved on one hand assigning data for census blocks to grid cells and on the other distributing data at the settlement level to appropriate grid cells.

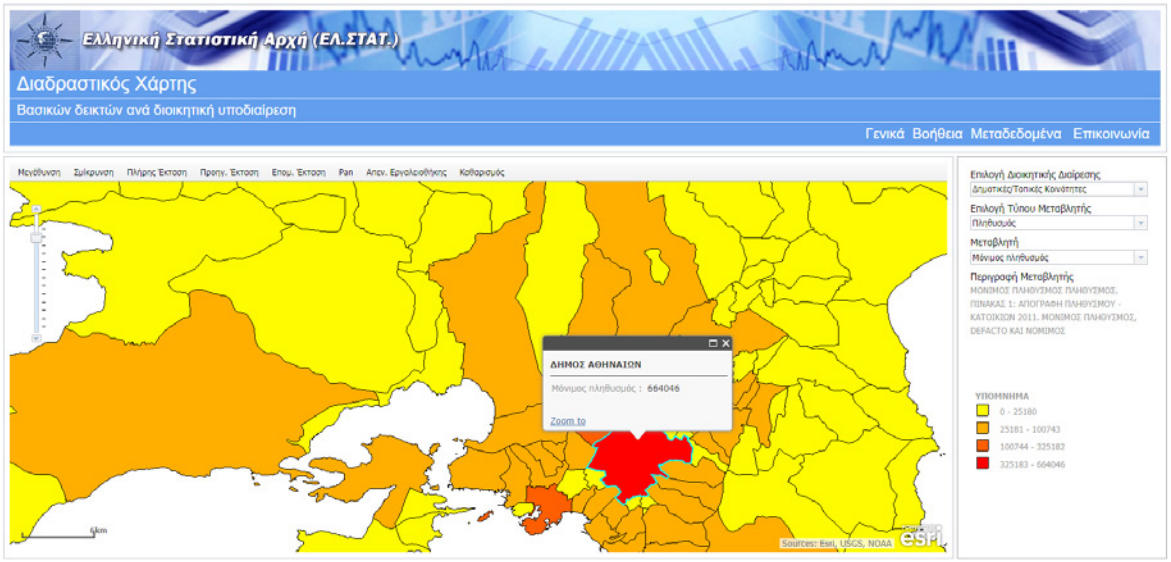
Figure 1: Example for the Attiki prefecture showing the 1 km^2 population grid results for 2001



- A correspondence table was compiled to match the codes used for the census block polygons and the codes used for the census blocks in the census data; these data were then matched to grid cells based on the centroids (geometrical centre) of each census block.
- For the distribution of settlement data the first step was to digitize the shapefiles forof the settlement's boundaries, using orthophotographs. The shapefiles were then overlaid with the grid layer and the census data for each settlement's resident population was assigned to the related grid cells in a proportionate manner assuming a homogenous distribution of people within each settlement.

Combining the results of these two actions resulted in a grid cell dataset which covered 92.6 % of the total resident population. The European Environment Agency's (EEA's) population density raster dataset (disaggregated with Corine Land Cover 2000) for the island of Crete were compared as a test case with the values calculated from this study and the differences were analysed.

Figure 2: Example of the web application for the area in/around the Greek capital of Attiki prefecture with results from the population census, by municipality/local community



A web mapping application was developed to map data for a wide range of administrative or statistical divisions of Greece, from regions down to local/municipal communities. A wide variety of statistical data were geo-referenced and entered into a database, including information on population, education, agriculture, construction, health and social protection, national accounts, tourism and the labour market. A geo-database was developed with spatial information linked to statistical data. Internationally recognised standards in the field of geospatial information — such as ISO standards and Open Geospatial Consortium standards — were adopted. The application is hosted and served from one of the Greek statistical office's servers.

RESULTS

Despite the deficiencies of the original data sets, a dataset for 1 km^2 grid cells was produced using the bottom-up estimation method. It is expected that a more accurate population dataset for grid cells can be produced from the 2011 census results which has a larger set of data referenced to census blocks. The web mapping application was designed to provide access to maps and to metadata as well as having a help function. A variety of tools were made available to users, for example to zoom in and out, to pan across the map as well as providing a previous and next area (similar to a back and forward feature of a browser). Users may select the administrative division they want to analyse, followed by a thematic category which then generates a list of available variables to select from. A map is then returned showing four classes (ranges) of data, along with a short text description and a legend defining the classes. Users can click on individual polygons to find precise values as well as the names of the selected administrative divisions.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!fJ73Hu
Hellenic Statistical Authority: http://www.statistics.gr/en/home/
Mapping portal: http://mapserver.statistics.gr/map/index.html

Hungary

Hungarian Central Statistical Office

Merging statistics and geospatial information in Member States; 2012 project; final report 18 March 2014

KEYWORDS: register, address, visualisation

PROBLEM

Within the statistical office there were 20 disparate and unrelated databases containing spatial data.

OBJECTIVES

The aim of the project was to connect this disparate set of statistical data and spatial data, combining information from the business register and population statistics through an address register and address directory to produce a geocoded data set to be used for creating maps.

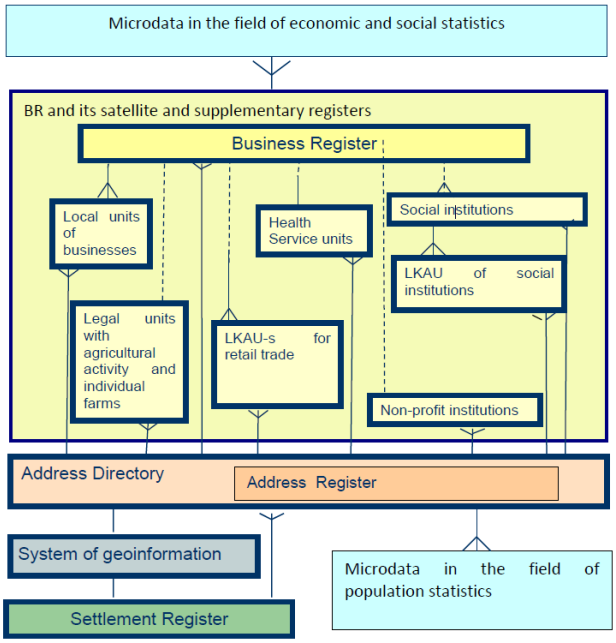
METHOD

The business register in the HCSO was derived from a variety of mainly administrative sources, including the Office of Government Issued Documents, the Registry Court, the tax office, the Treasury, or register questionnaires.

An address register existed within the Hungarian Central Statistical Office (HCSO) containing valid and approved addresses in a hierarchical structure and format: settlement, area, public space, real estate (house number, lot number), parcel (building, staircase, level, door number). However, its initial purpose was to support the conduct of population surveys rather than to have a complete register of the population. Alongside this, the HCSO also had an address directory which had other — non-approved — addresses. This information came in an unstructured and non-harmonised textual format from a variety of (mainly administrative) sources.

The central element of the geostatistics system developed by HCSO is an address directory, which identifies and manages addresses in the approved register, as well as addresses that have yet to be approved. The addresses used in business register and other satellite registers were adapted from textual addresses to a system of address identifiers that were linked to the address register or the address directory. These address identifiers are, in turn, linked to geocodes. By doing so, the HCSO has built-up a completely new system which provides the possibility to seamlessly merge geospatial and statistical information, by assigning geographical coordinates to addresses that are linked to statistical units.

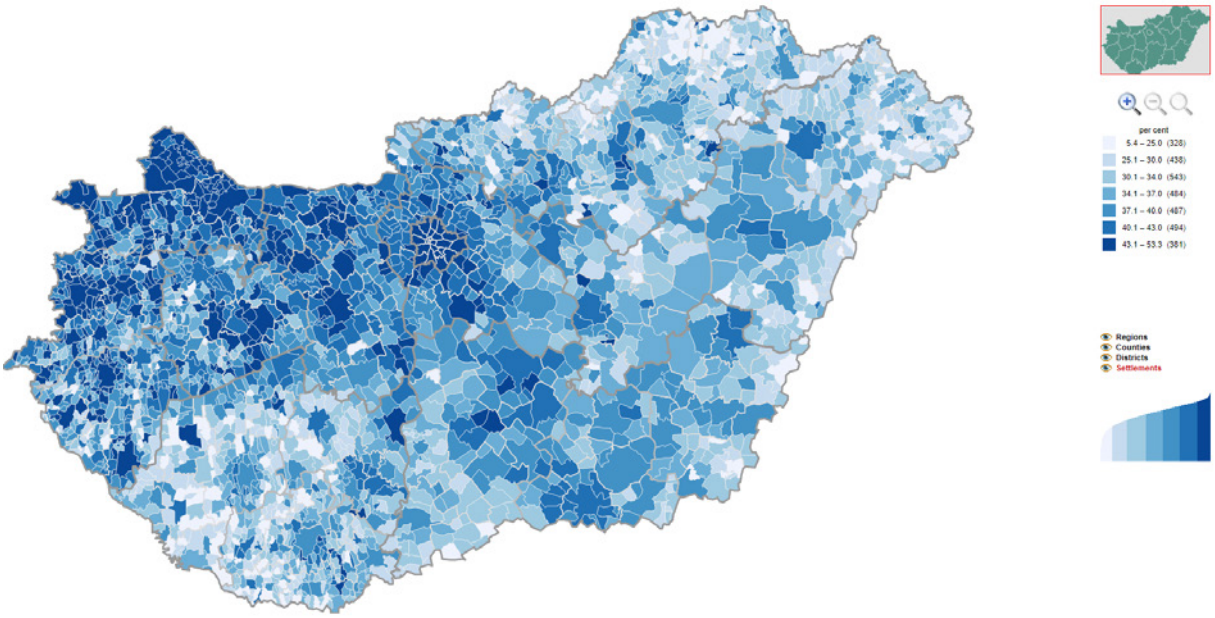
Figure 1: Example of the geostatistical system created within the Hungarian Central Statistical Office



Geocodes were available for the address list used for the census, with geocodes stored at the level of house numbers in the address register. Geocodes for new addresses and changes to addresses were bought from a Hungarian map company. For those addresses that were not approved, but stored in the address directory, a database function was created to identify the closest address from the address register.

This work allowed the HCSO to implement a uniform address management system across various registers, whereby the identity of addresses was established and a geographic code assigned. The functions of the address directory were modified so that it could manage any address coming from any register. This was done by: breaking

Figure 2: Example of the web interface developed by the HCSO showing results of the 2011 census detailing employed persons as a share of the resident population, by settlement



down addresses into standardised elements; devising systems to correct syntax, synonyms and errors; establishing connections to addresses already verified and included in the address register; devising a system to approve addresses not yet included in the address register.

At the end of the process, registers no longer contained addresses but rather address identifiers, to which the textual addresses were attached from an address book. This allows much easier control of addresses in terms of any maintenance and also means that a wide range of data may be visualised by drawing on the geostatistics from this harmonised system.

RESULTS

A geostatistical system has been developed linking information from business, population and social statistics to the address register or address directory, which in turn is linked through a system of geocodes to a map-making functionality.

A web interface was developed whereby users may choose from a number of broad statistical themes (such as population, social, economic or environmental statistics), then select one of the available indicators for that theme and one of the (up to four) levels of territorial typologies within the country (regions, counties, districts or settlements). The resulting map can then be personalised by selecting one of four criteria for determining the class boundaries and colour scheme. Users have standard navigation tools (zoom in/out and pan) as well as being able to mouse over any polygon of a statistical division in order to view the code and name of the area and its value for the selected indicator. For all except the most detailed level (settlements), clicking on a specific polygon generates a data tabulation for all areas at the same level of detail.

Furthermore, three sets of metadata were developed, one concerning the geocodes (source, data and accuracy), one concerning the source maps (name, source, date, accuracy, other characteristics) and one concerning thematic maps (source map, data source, date, method of creation, INSPIRE and ISO 19115 themes, terms of usage).

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!wM88Jf
Presentation: https://europa.eu/!Ub63Bp
Hungarian Central Statistical Office: http://www.ksh.hu/?lang=en
Mapping portal: http://www.ksh.hu/interactive_humaps?lang=en

Malta

National Statistics Office

STATAMAP: spatialisation and dissemination of statistics; 2012 project; final report 15 November 2014

KEYWORDS: visualisation, local administrative units

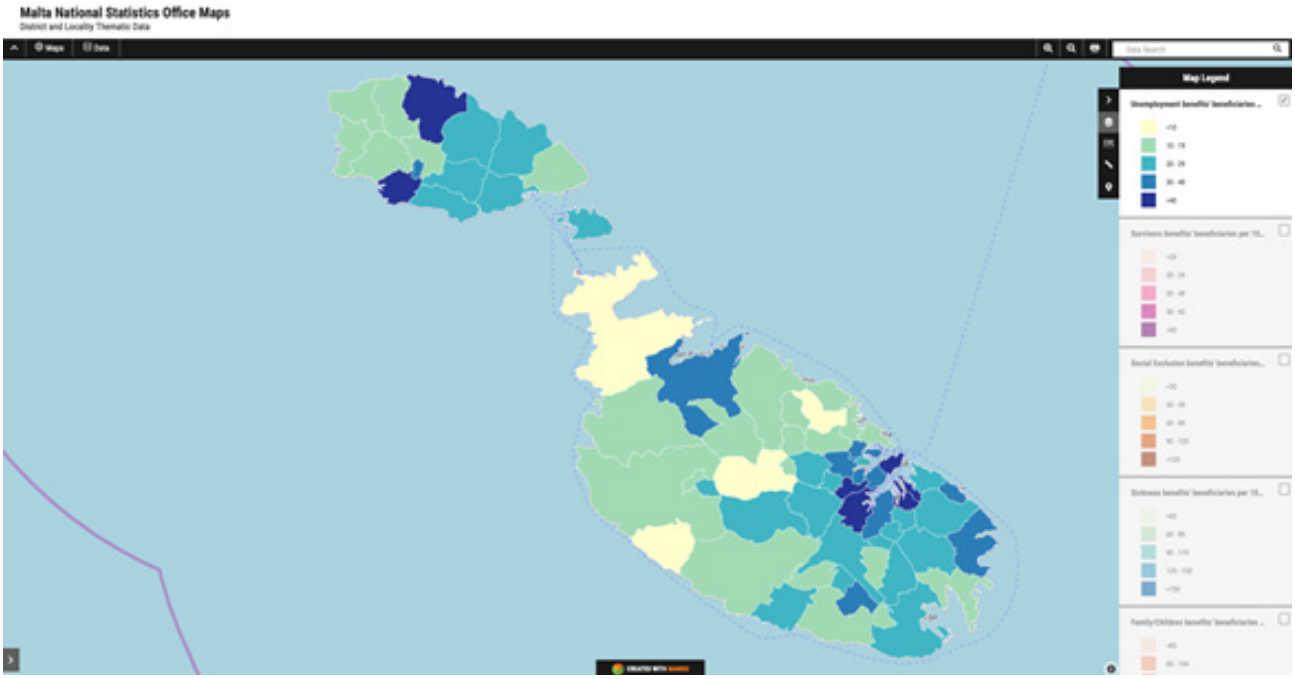
PROBLEM

The lack of a coherent strategy for developing georeferenced information, in terms of internal coordination and knowledge sharing within the national statistical office (NSO), in relation to other external data providers, and with respect to end users.

OBJECTIVES

The aim of this project was to review the NSO's capacity for geographic information system (GIS) spatial data creation, the generation of maps and the dissemination of data through its website. In other words, the project sought to create a working model covering the full data cycle from data sourcing through to dissemination which could subsequently be used by the NSO.

Figure 1: Malta National Statistics Office Maps



METHOD

The project started by examining the legislative and operational environment to understand how data were managed and disseminated as well as to identify any weaknesses. The project also looked at perceived needs of the general public and officials in terms of data dissemination.

The second phase of the project focused on the identification of datasets that could be used for mapping alongside the development and implementation of two training courses designed to demonstrate how to make use of these data.

The third phase concerned the development of dissemination technologies along with a training manual.

A network was set up between the national statistics office and other data providers in order to provide a means for the work to continue after the end of the project. It concluded by disseminating data and raising awareness of the results among local administrations, educational establishments as well as public and private organisations.

RESULTS

The project identified a set of principal datasets that were mainly taken from the 2011 census exercise. It created a transposition process, produced maps from the datasets and identified the best technology for online and media dissemination (based on employing INSPIRE data specifications, metadata and geoserver technologies). The focus of the project was on data for local administrative units (LAUs); at this spatial level there are 68 local councils in Malta.

The project also produced an analysis of (users') perceptions of spatial information as well as providing training to staff in the national statistics office.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!YJ76WV
National Statistical Office: https://nso.gov.mt/
STATAMAP: https://nso.gov.mt/en/Services/STATAMAP/Pages/STATAMAP.aspx

The Netherlands

Geonovum

Joining tabular and geographic data — merits and possibilities of the table joining service; 2012 project; final report 9 September 2013

KEYWORDS: table joining service, visualisation, merging data

PROBLEM

No institutional framework for linking statistics to locations (due to an absence of standardisation and interoperability).

OBJECTIVES

The general aim of this study was to explore the merits and possibilities of a table joining service (TJS), a geospatial standard developed by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/). A TJS offers a standardised web interface for automatically joining tabular statistical data to geographical information (administrative boundaries, postal codes, statistical units).

The study examined the concept and functional possibilities of using a TJS, existing TJS implementations and client applications. The implications, roles and tasks, as well as a cost-benefit analysis of the adoption and implementation of a TJS were also considered. As such, the main goal of the study was to gain an insight into the merits and possibilities of using a TJS for merging or linking tabular statistical data and geographical information.

METHOD

A TJS can be seen as an alternative method (to SQL and databases, GIS-based, WFS-based methods) in the context of the adoption and use of data services that are based on spatial infrastructures. A TJS has several potential advantages, such as its use in service-oriented architectures, as it is an open standard that promotes interoperability; it is also relatively simple and powerful, using the power of networked computers, while it is relatively easy to find public datasets through registries. At the same time, some potential disadvantages of TJS were assessed, such as the adoption of another service specification and another specific format for encoding data, or the lack of a mechanism within a TJS for uploading and creating tabular (or geographic) data.

A TJS can be considered as a supporting concept within spatial data infrastructures for joining data from various, distributed sources or infrastructure nodes. The concept of table joining in a distributed service-oriented architecture (over networks) requires that organisations implement their services according to a common spatial (identifier) framework. A common spatial and identifier framework consists of several aspects:

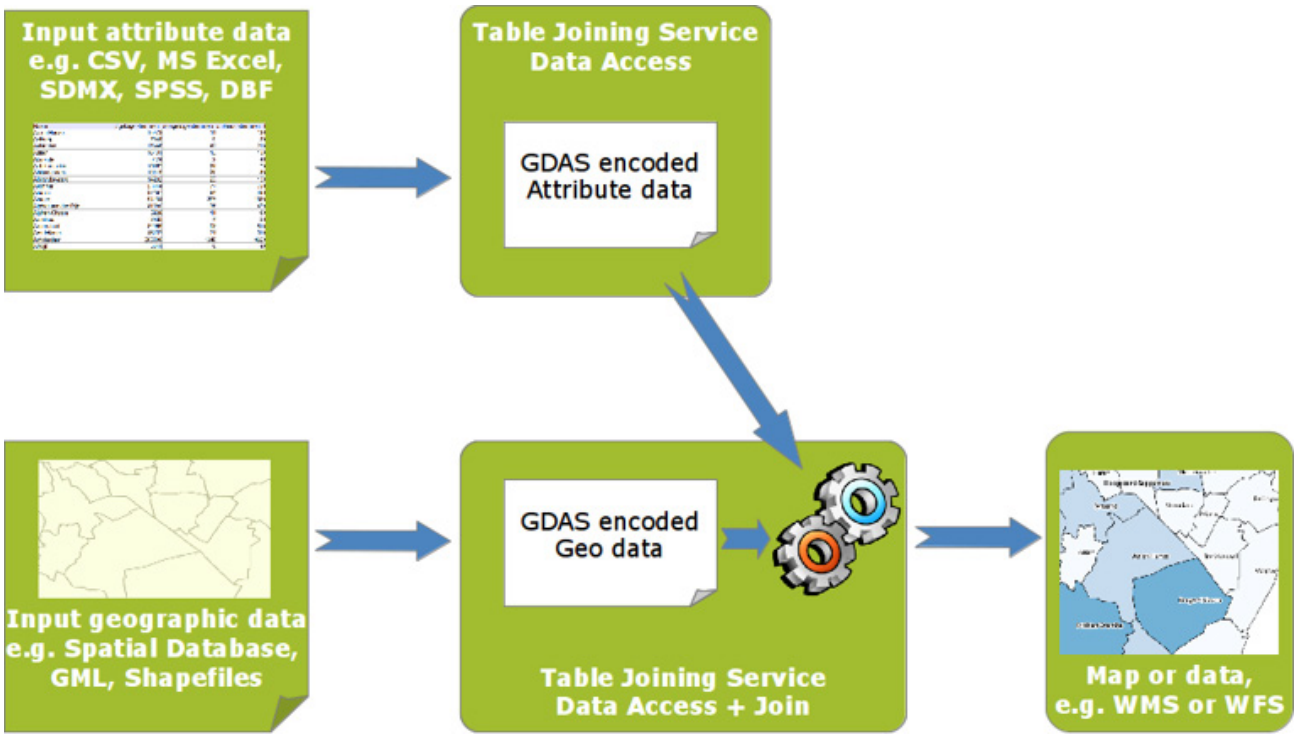
- data models for spatial frameworks based on a generic conceptual model;
- a generic approach for handling and encoding unique geographic identifiers;
- a registry for the maintenance of data models and identifiers.

The project conducted a cost-benefit analysis of using a TJS across a large number of public organisations in the Netherlands. These looked at a variety of costs, including the costs of software development (client and server), infrastructure set-up and maintenance, hosting and training; these were compared with the costs of manually joining datasets. Three scenarios were considered: the first assumed continuing to manually join datasets; the other two looked at a small and a large-scale implementation of a TJS.

RESULTS

The study concluded that a TJS has the potential to replace manual data joining operations in the daily data management practices that focus on thematic mapping and spatial statistics. The TJS standard offers a web-service interface to enable the automatic, service-oriented joining of tabular and geographic datasets across the web, while keeping original source data stored at the location of each individual data provider.

Figure 1: Input of data to create GDAS encoded data for Table Joining Services (TJS)



The cumulative cost of a small-scale implementation was estimated to be below the costs associated with manually joining datasets from the fifth year onwards, whereas for a larger scale implementation the cumulative costs of implementation were estimated to be lower by the eighth year.

Besides the monetary advantages, additional non-monetary benefits were also expected if a TJS was implemented:

- indirect efficiency benefits and higher quality output — the development of a framework requires a common agreement between many organisations offering spatial and tabular data, which may increase the quality and coherency of data management in several ways;
- societal benefits through better use of spatial statistics — more users would have access to spatial data through adopting a TJS solution, which would likely lead to an increased use of spatial statistical data with improved policy and decision-making;
- increasing and improving the quality of spatial statistical data exchange between organisations through the use of an international open standard — a TJS was expected to provide faster data delivery and a shorter 'time-to-market'; the use of open standards was also thought to have a positive impact on the independence of client software suppliers;
- benefits for organisations insofar as the use of TJS service-oriented architectures provide possibilities for monitoring data access and data use, thereby offering organisations the possibility to improve their services and service delivery.

The small number of TJS implementations found based on OGC standards and the lack of large vendors specialised in geographic information system (GIS) and TJS software implementations means that additional investment will be required before it is possible to test existing TJS software and client applications that make use of TJS.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!rV38vh
Total cost of ownership model: https://europa.eu/!xG44dv
Presentation: https://europa.eu/!gX97Gn
Geonovum: https://www.geonovum.nl
Statistics Netherlands: http://www.cbs.nl/en-GB/menu/organisatie/default.htm
Statline: https://opendata.cbs.nl/statline/#/CBS/nl/

Poland

Statistics Poland

Merging statistics and geospatial information in Member States, 2012 project; final report 10 February 2014

KEYWORDS: commuting, analytics, address, demographics

PROBLEM

The lack of spatial data beyond that traditionally presented for administrative divisions (gminas) led to growing calls from local government, scientists, entrepreneurs and individual consumers for the development of alternative statistics for a broad range of spatial divisions.

OBJECTIVES

This project set out to analyse the possibilities to present demographic data in area units other than administrative units that are currently used; it incorporated the manner of presenting statistical data in cadastral units, statistical units and grids. The project was designed to respond to (user) needs for providing data at levels lower than gminas (municipalities; the basic unit of territorial division in Poland), as expressed by local government representatives, scientists, entrepreneurs and individuals.

The project also aimed to develop spatial information for enterprise addresses and to combine this with demographic data to visualise commuting patterns.

Finally, the project looked at the possible use that might be made of spatial data for creating statistical indicators on land use to help the planning work of local government.

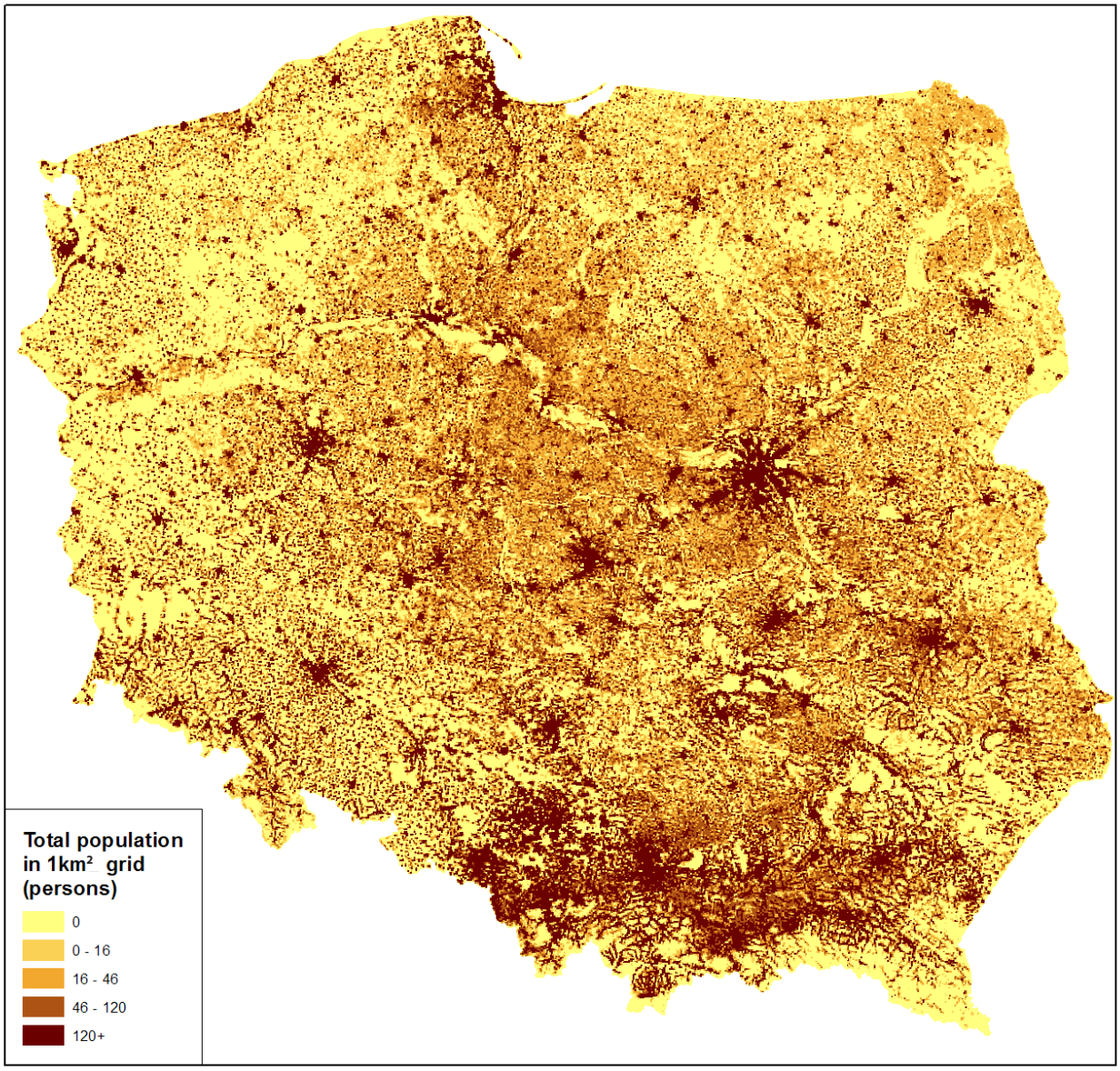
METHOD

The project combined data collected during the national agricultural census in 2010 with spatial data (for cadastral districts, statistical regions and census enumeration areas) from the national census of population and housing (conducted in 2011). As census data are collected with reference to address points, these data can be freely aggregated: the work involved an analysis of the manner of aggregation of demographic data to a number of other spatial classifications. Specifically this was done to produce data for statistical units (statistical regions and census enumeration areas), cadastral units (geodesic precincts) and for a 1 km^2 grid, although the latter could be modified to reflect any less detailed division of space.

A database of address points representing the location of enterprises was created. Several sources of enterprise addresses were identified (data from the social security system, the Ministry of Finance, the tax authorities, and the statistical business register) and the names and identification numbers of gminas (municipalities) were corrected using the statistical office's register of territorial divisions. Workplace coordinates were assigned to people based on the address identification system of streets, real estate, buildings and dwellings which is part of the statistical office's register of territorial divisions, as well as using the statistical business register and data from the database of topographic objects (of the national mapping agency). In cases where a match could not be achieved between the address identification system and the addresses given in the source for enterprises, a number of simplifications were made in order to try to use as much of the address information as possible.

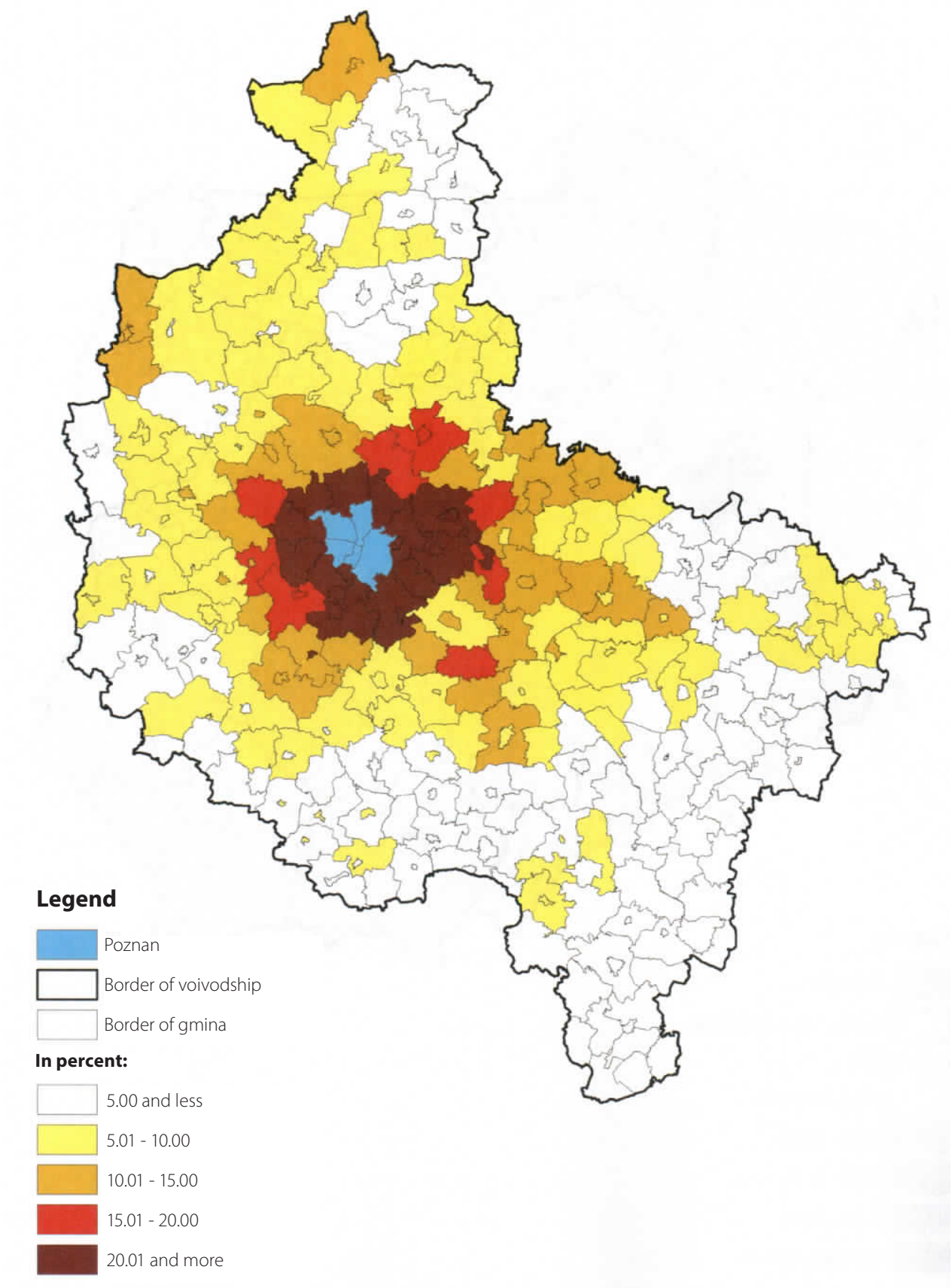
The Centre for Urban Statistics of the statistical office in Poznan developed a methodology for surveying commuting with the use of data acquired for the purposes of the 2011 census and these showed the possibilities for linking geo-information and statistical information. For the needs of censuses, a spatial database of address points representing the location of residential buildings in Poland was created. Combining this with the new database of address points for enterprises, it was then possible to map commuting statistics. A dataset was prepared containing records for people who had reported in the census that they commuted: the dataset included information on an individual's age and sex, as well as various identifiers for their place of work including the XY coordinates, various identifiers for their place of residence including the XY coordinates, and the type of work performed. From this information the (direct) distance between the XY coordinates of the place of work and the place of residence could be calculated. Two datasets were produced, one focusing on commuter departures (from home) and the other on commuter arrivals (at work).

Figure 1: Example of spatial visualisation of demographic data, Poland



In addition, the Regional and Environmental Surveys Department conducted an analysis of the possible use of spatial data for creating statistical indicators on land use planning. This involved a conceptual stage (for example, defining survey methods), data collection, data evaluation and then data analysis. For one voivodship (the highest level of administrative unit in Poland that corresponds to a province in several other countries), a database of topographic objects was used, along with an orthophotograph (an image that has been geometrically corrected to give a uniform scale, similar to that in a map); the data were also compared with the land and building register. The data in the database of topographic objects were evaluated to look for errors. This was done by checking the information available for a sample (5 %) of 1 km^2 grid cells to see if the database correctly contained the objects observed on the orthophotograph. Around four fifths of the grid cells had the same information in the database as observable on the map for roads, with this share rising to 100 % in rural areas; most of the differences in urban areas concerned short sections of roads for groups of new buildings. Concerning buildings, the database was at least 75 % compliant for more than four fifths of all grid cells. Again, grid cells in urban regions were more likely to show higher non-compliance with the database, often related to recent or on-going construction activity.

Figure 2: **The share of persons arriving to work in Poznań in the number of employees in the gmina of residence in 2011**
(in voidvoship scale)



RESULTS

After performing disclosure control on the census data, thematic choropleth maps were prepared and published on a platform developed for the spatial visualisation of statistical data — the Geostatistics Portal. This resulted in the presentation of demographic data that was divided in a different manner to traditional administrative divisions. There was a strong level of demand for the resulting information, especially for demographic information at detailed levels, below that of gminas (municipalities).

The work on assigning address points representing the location of enterprises resulted in a complete dataset with XY coordinates for the place of work for all employed persons, either based on an exact identification of the address of an enterprise or an approximation thereof. The database may be used to visualise many phenomena and statistical data that are related to surveys conducted at the place of work.

Various analyses of these commuter datasets were performed, for example, simply identifying how many people in each voivodship (of which there are 16 in Poland) were commuters, showing at a very detailed level where people worked, or showing the share of commuters among all persons employed. As well as presenting these data according to various administrative and statistical areas, analyses were also performed for 1 km^2 grid cells, showing where people commuted from and where the numbers of net inward commuters was highest. Further analyses were performed focusing on commuting patterns to particular locations, for example showing the origin of people commuting to a specific city, as well as analysing the structure of commuter flows by age or by sex, and the average distance travelled between a commuter's place of residence and place of work.

The work on land use evaluated the quality of the database of topographic objects as good. The gaps in data were generally in heavily developed areas where new infrastructure emerged and new buildings were constructed (suggesting that identification issues were more likely linked to new building work rather than any to any underlying issue in relation to the methodology for identifying/registering objects). This subproject also developed methods for acquiring information on the density of buildings and road networks, starting from the number of buildings, the extent of the built-up area and the length of roads; this was done for gminas (municipalities) and for 1 km^2 grid cells.

The conclusions drawn from this work confirmed that it is possible to produce data: i) at a detailed level (for 1 km^2 grid cells) for buildings which previously were only available for gminas and ii) on road infrastructure which was previously not even available for gminas.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!PF87BN
Annex I to final report: https://europa.eu/!Yw86uD
Annex II to final report: https://europa.eu/!cH68kP
Statistics Poland: http://stat.gov.pl/en/
Commuting to work in Poland — results of the national census of population and housing 2011: http://stat.gov.pl/en/regional-statistics/regional-surveys/commuting-to-work-in-poland--results-of-the-national-census-of-population-and-housing-2011/
Mapping portal: http://geo.stat.gov.pl/en/
TERYT register: http://eteryt.stat.gov.pl/eTeryt/english.aspx

Slovenia

Republic of Slovenia
Statistical Office

REPUBLIC OF SLOVENIA
STATISTICAL OFFICE

Merging statistics and geospatial information in Member States, 2012 project; final report January 2015

KEYWORDS: dissemination, address, registers, grid

PROBLEM

The Statistical Office of the Republic of Slovenia (SURS) managed a large number of datasets which were disseminated over a variety of different platforms (including cartographic presentations) with the lack of any coherent and integrated system for the dissemination of geospatial statistics.

OBJECTIVES

The objectives of this project were to:

- improve the integration of geo-information and geo-referencing within the statistical production process;
- illustrate how linking geo- and statistical information provides additional value and creates new information;
- design innovative web applications to show the spatial distribution of statistics.

METHOD

The statistical office worked with the Geodetic Institute of Slovenia, which provided access to information technology (IT) staff with knowledge in software development to develop an application. The institute is the leading Slovene public institution for geodetic, cartographic, geo-informatic and hydrographic research and development and is a part of the national geodetic service. Furthermore, it already had experience implementing INSPIRE.

The Register of Spatial Units is the main source of administrative geospatial data in Slovenia. It contains over 40 administrative and non-administrative official territorial divisions of Slovenia and is currently managed by the Surveying and Mapping Authority of the Republic of Slovenia (GURS). In 2008, a project introduced square grid vector layers with seven basic grids (grids ranged from 100 m * 100 m to 10 km * 10 km).

As a part of this project, the statistical office reviewed the statistical data that were already available through interactive cartographic visualisations, in order to identify: whether the data were relevant; whether other data should/could be added at least at a regional level; whether some data could be presented for grids; whether there were confidentiality issues.

Having identified those data sets that could be geocoded a number of sources were identified together with areas where they could be improved. For example, information on motor vehicle ownership in the Central Register of Motor Vehicles and Trailers was linked to the Central Population Register or the Business Register in order to have more accurate ownership addresses, before disseminating these data. Other datasets that were assessed included the Statistical Register of Employment and various datasets related to income. A number of organisations producing official statistics outside of the statistical office were also involved, for example the National Institute of Public Health which had data on standardised mortality rates and information for physicians.

Although all statistical data were either geocoded or geo-referenced throughout the entire statistical production process, the Slovenian statistical office did not have a permanent solution regarding the production of geospatial statistical data for dissemination purposes. The project addressed this issue through the design and development of a web application — STAGE (Statistics & Geography) — as a central dissemination tool for geospatial statistics. As well as bringing together statistical data, the development of STAGE was an opportunity to implement INSPIRE's metadata rules. STAGE replaced two Flash based applications (ISAS and KASPeR).

The project also considered confidentiality: prior to the 2011 census, the standard geospatial units for the dissemination of official statistics were administrative units — the smallest being settlements; grids were produced only upon request and each request was discussed by the Data Protection Committee. The statistical office performed many analyses in order to define the confidentiality rules that should be applied to grid statistics in Slovenia. It was decided that information on education and economic activity should be subject to disclosure control, whereas other information could be published without any controls for disclosure, such as the size of the population by age group and sex, the number of households and dwellings by size or the number of buildings by year of construction.

RESULTS

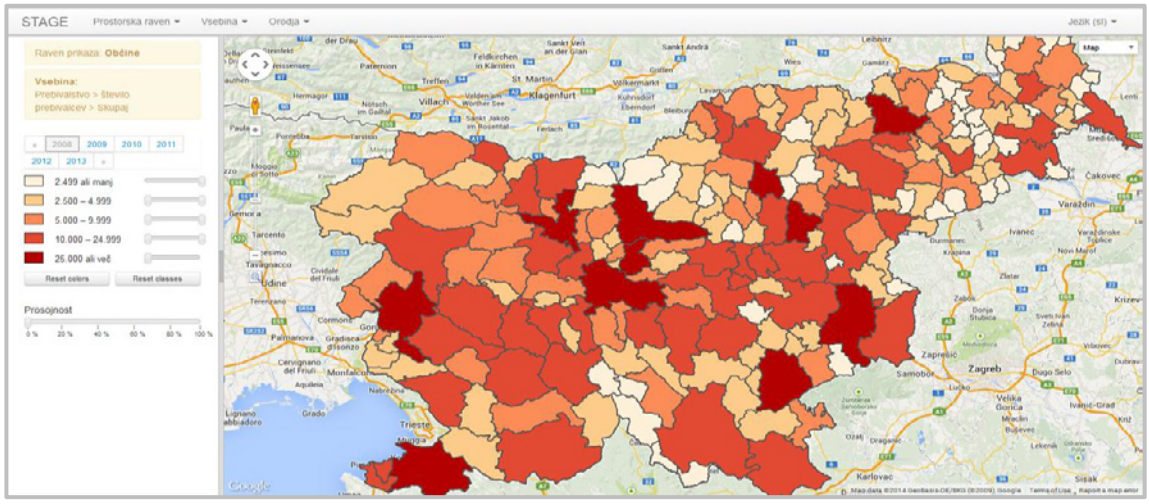
This project resulted in the development of an integrated system for the dissemination of geospatial statistical data in STAGE. STAGE is an interactive web application for cartographic visualisation of geospatial statistics and downloading data: it displays data in the form of choropleth maps. It is designed as an INSPIRE compliant application. The merging of geospatial information and geocoded statistical data was significantly simplified through the development of STAGE. Official geospatial statistical data can now be integrated into the data infrastructures used by other major data providers or simply combined with user's own geospatial datasets to create additional added value. A list of variables was identified for which statistical data are now published on a regular basis in STAGE.

The implementation of STAGE opened a new dimension for the dissemination of official statistics in Slovenia. Many public agencies are following the idea of presenting their data through grid-based maps.

The quality of data in some of the source databases was also improved through this project when checking data prior to its use in STAGE.

Awareness of the importance of the spatial dimension of statistical data was also increased within the national statistical office.

Figure 1: Example of STAGE portal showing population in the municipalities of Slovenia



FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Pb64Kj
Geodetic Institute of Slovenia: http://www.gis.si/en
Republic of Slovenia Statistical Office: https://www.stat.si/StatWeb/en
STAGE (mapping) portal: http://gis.stat.si/
SI-STAT database: http://pxweb.stat.si/pxweb/dialog/statfile1.asp
Statistical data in thematic cartography: https://www.stat.si/TematskaKartografija/Default.aspx?lang=eng
Slovenian statistical regions and municipalities in numbers: https://www.stat.si/obcine/en

Slovakia

Statistical office of the Slovak Republic

ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY

Representing census data in a European population grid, 2012; undated final report

KEYWORDS: demographics, census, address, grid

PROBLEM

Development work on representing census data on the population grid was incomplete, as address points needed to be consolidated and then linked to statistical data from the census.

OBJECTIVES

This project was a continuation of a previous project and concerned the development of datasets based on a population grid. The aim was to connect the results of the 2011 population and housing census (SODB 2011) with a population grid to create a dataset for the latter — according to GEOSTAT methodology — with statistics on the total population by sex and by age.

METHOD

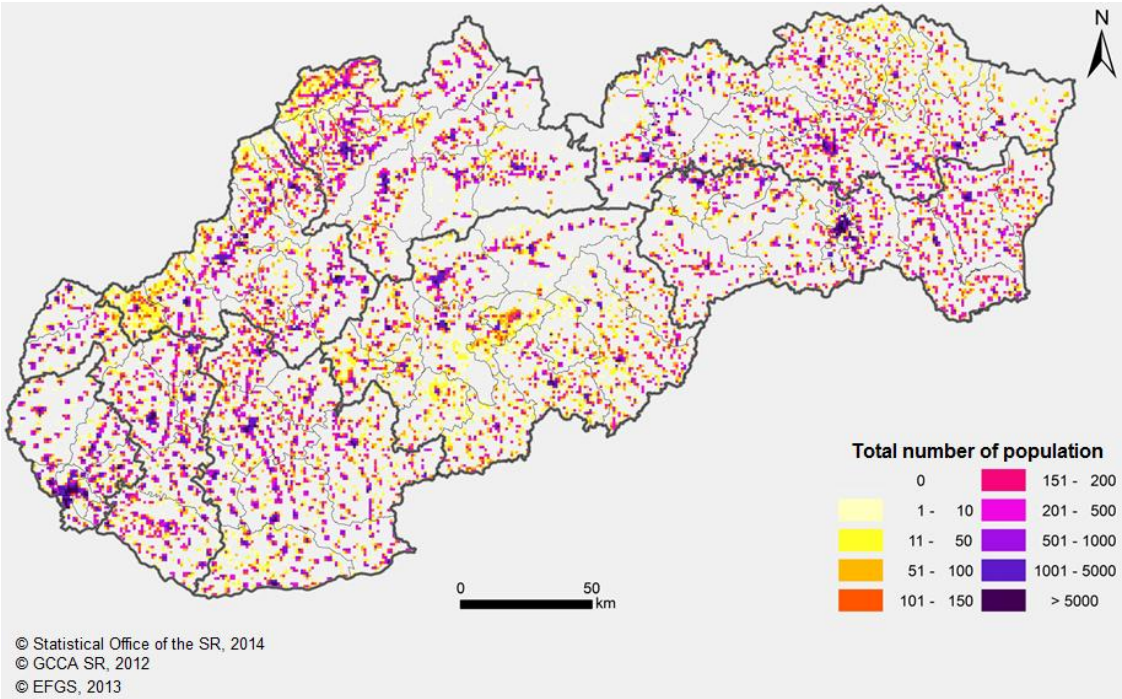
As part of the 2011 census exercise an Oracle database was created. Boundary information was received from the Geodesy, Cartography and Cadastre Authority of the Slovak Republic, while data for residential units were provided by the Slovak Environmental Agency. Information from the population registry from the Ministry of Interior of the Slovak Republic was linked to individual address points. An application was developed to facilitate the creation of new address points or to modify or delete existing address points during the data collection exercise for the 2011 census.

RESULTS

The result of using this application to add, delete or modify address points was the establishment of a complete layer of address points for the whole of Slovakia, combining a range of geographical information (such as administrative and census districts, postal addresses, XY coordinates) with information about inhabitants, in other words a geo-based dataset for the 2011 census.

From this dataset a range of grid-based statistics was developed containing information on the total population in each grid cell as well as an analysis of demographic information by sex and by age.

Figure 1: The total number of people in the 1km x 1km raster, Census 2011



While this dataset for grid-based statistics was available for a relatively small range of indicators that were regarded as non-confidential, this was not true for local and regional statistics. Indeed, methods were developed in relation to data presentation and the protection of confidential data, which involved a one-off implementation of a method known as targeted record swapping. The data for individuals were assessed against a number of criteria, resulting in a risk score. A swap rate was calculated for each NUTS level 3 region and from this the number of households to be swapped was calculated for each municipality within that region. Within each municipality a set of households was selected proportional to their risk score. For selected households, addresses in the database were swapped with another household with a similar age-sex structure within the same district (which is a local level situated between a municipality and the NUTS level 3 region). As only the address is swapped the relationship between any other variables in the database are not disturbed, although there might be an impact on aggregated results for some municipalities. An advantage of this approach is that the whole database can be treated once, ensuring that any tabulations made from the database — whether standard or non-standard — are consistent with each other.

The final initiative was to develop a web mapping service for users, incorporating the geo-dataset.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Kw73PM
Statistical office of the Slovak Republic: http://slovak.statistics.sk
My municipality in statistics: https://slovak.statistics.sk/wps/portal/ext/products/municipality/
Register of territorial units: https://slovak.statistics.sk/wps/portal/ext/Databases/REGPJ/

United Kingdom

Office for National Statistics

Office for National Statistics

The development of a web application for the semantic visualisation of geostatistics, 2012 project; final report January 2015

KEYWORDS: linked open data, visualisation, URI

PROBLEM

Statistics are measures of things, so every statistic has a geographic value. The Office for National Statistics (ONS) decided to investigate the potential use of 'linked data' as a mechanism for publishing statistical geographies and linking them to statistical data.

OBJECTIVES

Formally, this project had five key objectives, namely:

- to establish what difficulties existed with publishing statistical geographical information (for locations or areas) as geo-linked data;
- to make recommendations on best practices for publishing statistical geographies as linked data;
- to set-up a system for disseminating statistical geographies as linked data;
- to provide value added functionality through the development of a human-readable interface to the linked data;
- to investigate the potential issues with publishing statistics that link to statistical geographical data to other types of related linked information.

METHOD

The provision of technical infrastructure was outsourced using an existing platform called OpenUp, which provides a Resource Description Framework (RDF) — a specification for metadata.

In the United Kingdom there was already a single unique identifier for each geography, known as a GSS code, comprising nine digits. An early step in the project was to use these codes to develop a structure for unique resource identifiers (URIs) that are HTTP encoded. As part of the guidelines of the UK Location Programme it was decided to use data.gov.uk as the domain. For spatial things (real-world phenomena that have a spatial extent or position) the word statistics was added as a theme while for spatial objects (abstractions of spatial things) the word location was added as a theme.

For the development of URIs for statistical geography on spatial things, the following structure was adopted:

http encoding	theme	domain	spatial type	class	identifier
http://	statistics.	data.gov.uk	id/	statistical-geography	{GSS code}

For spatial objects the structure was linked to INSPIRE requirements and so included an element for the INSPIRE theme for statistical units. It also included geometry as a class to distinguish these objects from points or lines. This resulted in the following structure:

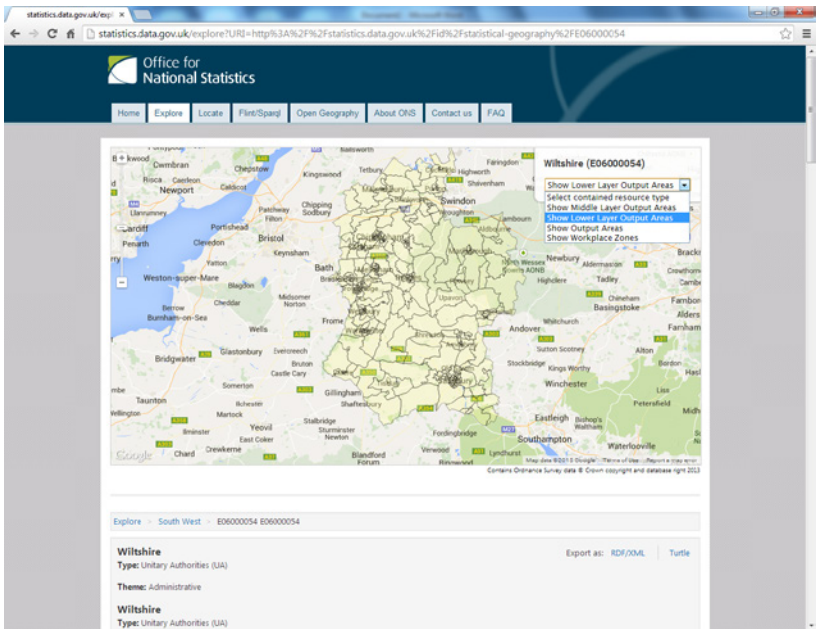
http encoding	theme	domain	spatial type	INSPIRE theme	class	identifier	resolution
http://	statistics.	data.gov.uk	so/	su/	geometry/	{GSS code}	{resolution code}

Having developed a structure for the URIs the main work was to consolidate data from a range of ONS products — using geography as a common linking element — in order to publish data online using a format based on a resource description framework (RDF). The preparatory work involved identifying which datasets would/could be converted alongside creating a data model for each one (using TopBraid Composer), which in turn led to a specific data vocabulary to describe all of the elements of the dataset in question. Scripts (in Python, which is open source) were then developed to use the data vocabularies and models to convert data from a basic format (*.csv files) to the RDF format. The resulting data were then loaded through an open data portal to the OpenUp platform, from which they were made available to end-users. The final step was to put a set of tools in place so that people (as opposed to machines) could find the information they sought.

ONS branding was added to an existing interface that could bridge the gap between the linked data and the query language (SPARQL): effectively the interface provided human readable tools for generating queries in SPARQL (without the user having to understand SPARQL) and then exported data in various formats such as *.xml or *.csv.

A visualisation tool called the Explore tool was also developed. This displayed selected geographies on a base mapping layer. Users could enter a geographic name or postcode (or select names from a list) and then be provided with all of the information available for that geography. A Locate tool was also developed. With this, users could identify an area of a map (a box with a specific boundary) in the form of a rectangle or freehand polygon with information for various geographies within the selected area.

Figure 1. Drop-down menu for selecting a nested hierarchy



RESULTS

A URI was developed for the NSOs geostatistics.

The various products that were published were: a Code History Database, the Register of Geographic Codes, the National Statistics Postcode Lookup and geographical boundaries.

In addition, user tools were developed facilitating not just automatic (machine) interaction with the data, but also facilitating data use by individual people.

Once completed, the system moved from being a special project to being integrated within the normal working environment of the NSO which, among other things, involved developing a sustainable maintenance process and developing skills within the organisation.

Having completed this stage of the development work, the focus subsequently moved on to linking more statistics to the geographic data so that a wider range of statistics was made available for any particular geography. The statistics disseminated through the tools developed for linking data were also disseminated through traditional products. In the future these tools might be discontinued as the data for different geographies migrate to a single data dissemination platform.

Linked data is a feasible format for delivering the requirements of INSPIRE. Providing data in RDF format makes it possible for data from several disparate sources to be connected together and delivered through a single user application.

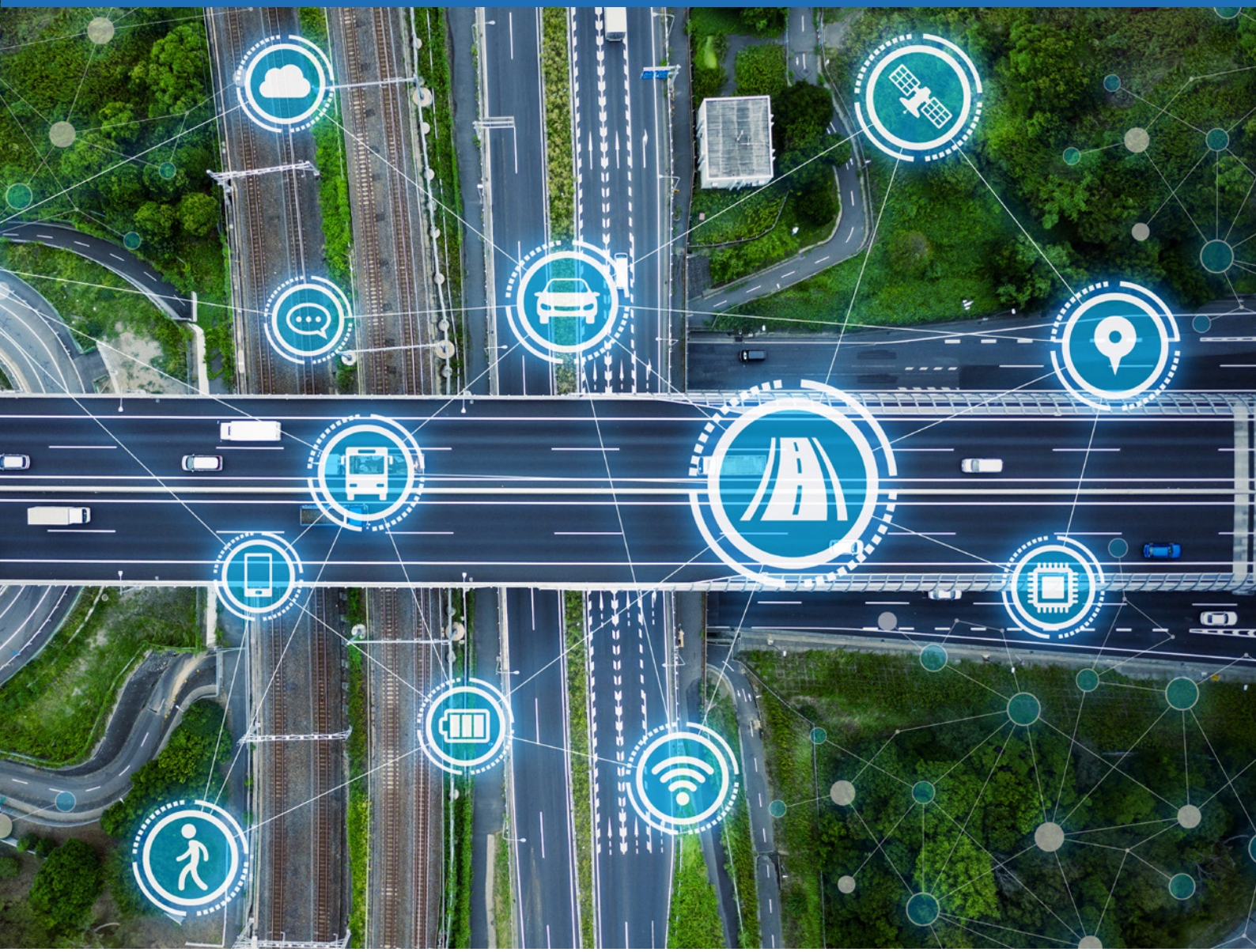
FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!tv96Vr
Office for National Statistics (ONS): http://www.ons.gov.uk/
ONS geography linked data site: http://statistics.data.gov.uk/
ONS open geography portal: http://geoportal.statistics.gov.uk/
Use of open standards:

- SPARQL 1.1 (http://www.w3.org/TR/sparql11-query/)
- Turtle — terse RDF triple language (http://www.w3.org/TeamSubmission/turtle/)
- RDF (http://www.w3.org/TR/rdf-primer/)
- URI sets for the public sector (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf)

B

2013 projects



Bulgaria

National Statistical Institute

НАЦИОНАЛЕН СТАТИСТИЧЕСКИ ИНСТИТУТ

Merging health care statistics with spatial information; 2013 project; final report 8 January 2016

KEYWORDS: health care, register, localisation

PROBLEM

There are considerable differences between different regions within Bulgaria as regards the availability of health care services. The national statistical office therefore sought to establish a spatial component within its production process for health statistics in order to help the Ministry of Health ensure that each region could be in a position to provide sufficient access to a broad range of medical facilities.

OBJECTIVES

The aim of the project was to improve the statistical production process through the integration of spatial information on core health care statistics and register-based data maintained by national statistical bodies in Bulgaria, providing more extensive statistical data for analysing the health of the population and the accessibility to health care services at a various territorial levels.

METHOD

The project focused on the localisation of health care statistics by using geographical location descriptions from administrative sources to analyse regional differences in health care services and capacity. In a first step, an iterative process was used to select a range of indicators relevant for an analysis of regional health care patterns. The indicators selected covered the following domains:

- demographic change (total population, population by age group, natural increase);
- health care services (establishments, available beds, physicians);
- health outcomes (births, deaths and causes of death).

The next step was to collect data from administrative registers, including those for in-patient and out-patient establishments/hospices, in particular collecting identifiers such as the name of each establishment, its address, municipality and district. Information was also collected from the cadastral register, such as the functional use of buildings and coordinate references.

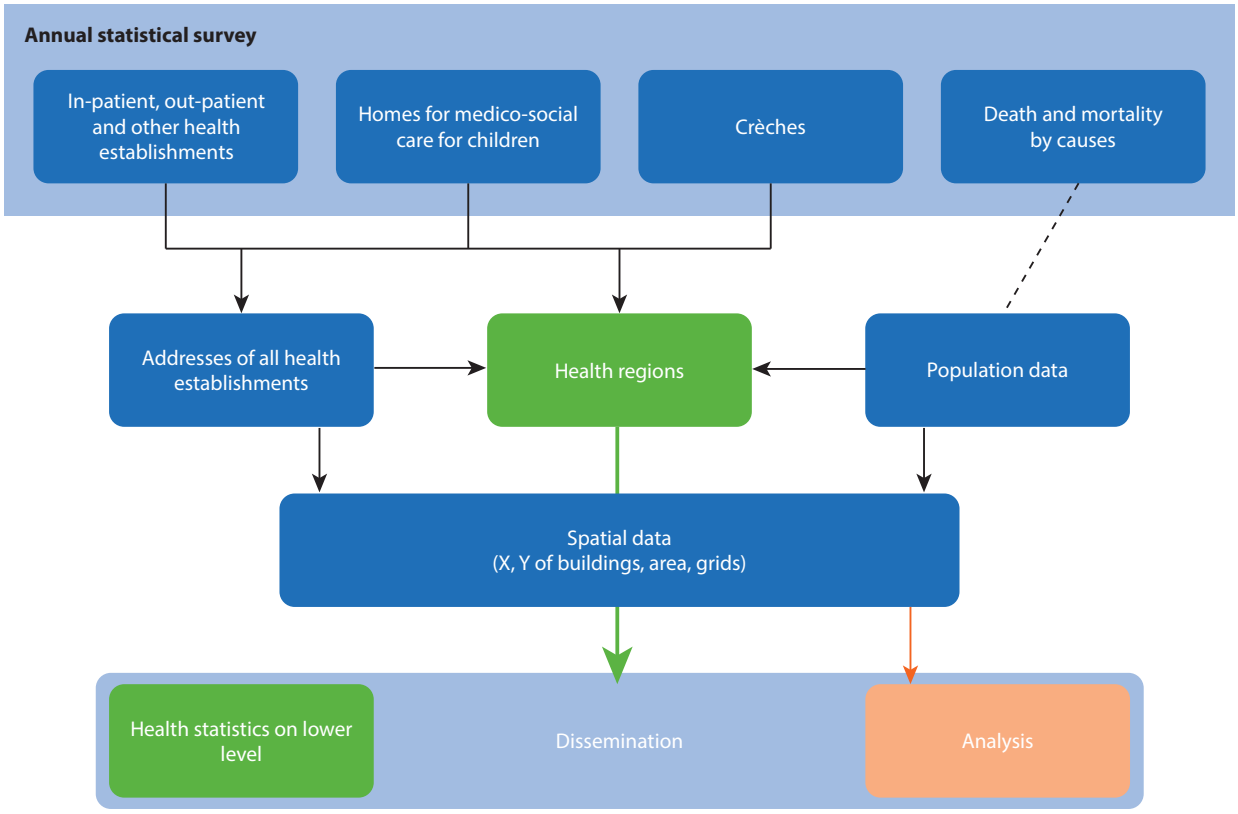
Each health care establishment was subsequently geo-referenced to the address of the service location (other than for a small number of out-patient establishments that were instead referenced to their postal registration). The address information was used to link register data with spatial data from the National Geodetic, Cartography and Cadastre Agency (NGCCA). In those cases where the cadastral information was scarce, street view was used to accurately position the location of health care services. Information was subsequently aggregated from local administrative unit (LAU2) geometries to health region geometries.

One of the main challenges was to harmonise the address information from health care and cadastral registers, although experience gained from previous exercises for the census proved invaluable. Furthermore, future updates were expected to be less of a burden given there is only a relatively small number of changes in the stock of health care establishments from one year to the next.

RESULTS

The results from this project have led to a new set of geo-referenced health and population data with information on health care establishments linked to data from statistical surveys. These results facilitate a more detailed analysis of accessibility issues in relation to health care services which provides the Ministry of Health with an opportunity to assess ways of improving public health planning at local, regional and national levels.

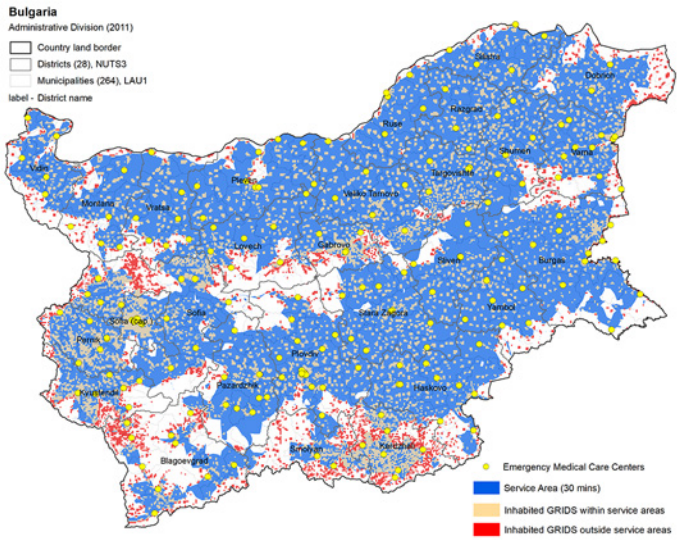
Figure 1: Organisation of the project



In the future it is possible that the work carried out during this project could be extended to cover spatial analyses of the distribution of patients having used health care services, or analyses of emergency (112) calls, with the goal of improving the spatial distribution of emergency teams to ensure a rapid response for patients.

Furthermore, it is possible that the experience gained during this project may be put to use when implementing similar projects for other statistical domains, outside of the area of health statistics.

Figure 2: Service areas within 30 minutes driving time of an emergency medical care centre



Source: Statistics Bulgaria

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!tJ98yu
National Statistical Institute: http://www.nsi.bg/en

Germany

DESTATIS — Federal Statistical Office

Merging statistics and geospatial information — the urban contribution to the European spatial data infrastructure; 2013 project; final report December 2015

KEYWORDS: urban audit, address, visualisation

PROBLEM

Post geo-coding of information was considered an expensive and a time-consuming exercise that was disjointed insofar as similar tasks were being repeated for individual registers.

OBJECTIVES

The KOSIS-Association (municipal statistical information system) is a network formed by 125 German cities (as of the end of 2013) authorised to produce Urban Audit statistics on cities; this data was supplemented by information from national and federal statistical offices, as well as other sources. KOSIS-Association led the project, known as the Merging Project, which sought to develop methods for harmonising information on German cities so that it followed European Union spatial data infrastructure rules (INSPIRE). The project sought to rectify several issues linked to the geo-coding of data, as well as identifying and establishing best practices for the development of administrative registers for small-scale statistics. By doing so, it sought to bridge the gap between municipalities (as providers of addresses) and administrative registers (as users of addresses) through the creation of a central address register that conformed to INSPIRE rules.

The project had the objective to:

- improve the integration of spatial information and geo-references into statistical production processes, particularly for administrative registers at the municipal level;
- establish processes that would allow geo-referenced information to be harmonised and continuously updated;
- illustrate how geo-referenced statistical information and its corresponding metadata could be used to create additional value, for example through web applications to show the spatial distribution of a wide range of socioeconomic statistics.

METHOD

One of the first tasks undertaken was to compare different methods used to collect information on addresses and spatial statistical units: this was based on contrasting the traditional method of collection used in Germany with that applied in the Netherlands.

Municipalities were found to be the principal source of information for addresses: they were responsible for the creation and maintenance of house numbers and street names (locators within INSPIRE). In Germany, there were many different administrative registers maintained in isolation by different departments of municipal administrations. By contrast, in the Netherlands all such registers were linked through a central address register

which provided updates automatically when new or revised information was encoded.

A simplistic model of the situation in Germany is shown in Figure 1, whereby geo-referencing was conducted in a sequential, unrelated manner. An alternative presentation, based on a simplistic model of the situation in the Netherlands is provided in Figure 2, where geo-referencing was an integral part of the central address register.

A proposal was made as to how addresses and spatial statistical units could be recorded with descriptions of the required data elements and processes, as well as details of how these could feed into city, state, regional, national and European infrastructures, with particular emphasis on geo-coding data in administrative registers. The review considered the provision of a central address registry as beneficial for improving the quality of geo-coded administrative data, providing an opportunity to:

- assign and store information in a single step;
- determine a unique location for each entity (rather than having the potential for different geo-references across unconnected registers) with its coordinates;
- improve the quality and homogeneity of addresses;
- access up-to-date information for house numbers and street lists;
- provide automatically details of address changes to other registers;
- allow for on-the-fly calculations, such as joining data from different registers or aggregating data to different territorial typologies;
- make considerable efficiency and cost savings.

Figure 1: Unrelated registers with no relationships for managing addresses

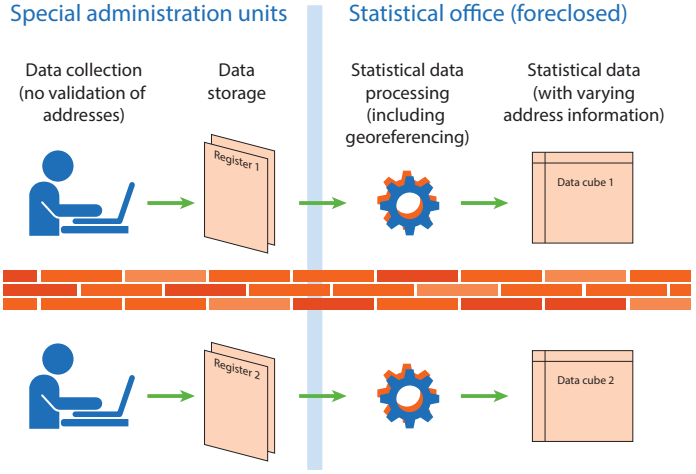
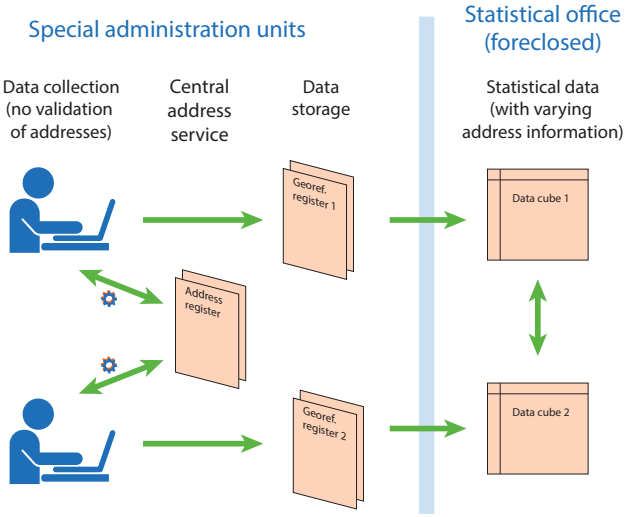


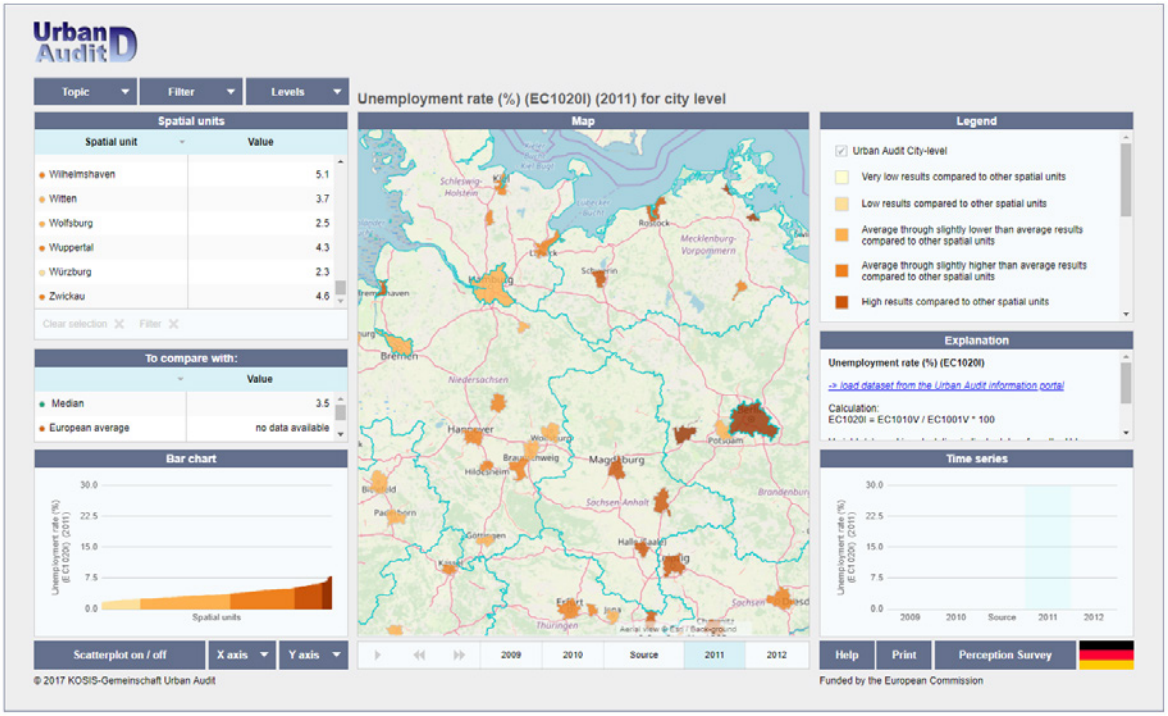
Figure 2: A central address register with in-built geo-referencing



The next task considered by the project was to develop a proposal for a definition that might be applied for comparable intra-city spatial units. On the basis of discussions, workshops and reports, the approach recommended was for intra-city observations and analyses to be conducted for sub-city units of approximately 5 000 inhabitants (large enough to guarantee data protection, but small enough to allow for meaningful analyses). In those cities which did not have or were not in a position to create such territorial units, grid cells were considered a good alternative as they offered the following advantages: they required little maintenance; they were stable over time; they enabled observations of certain phenomenon across municipal boundaries (for example, if monitoring segregation or gentrification).

The penultimate action for the project was to illustrate how linking geo-information and statistical information and corresponding metadata could provide additional value. This was done through highlighting the role that a DUVA-interface may play, enhancing standardisation through the use of data descriptions for common data sources to facilitate data exchange, while providing a central instrument that may be used to manage metadata (its description, capture, processing and presentation).

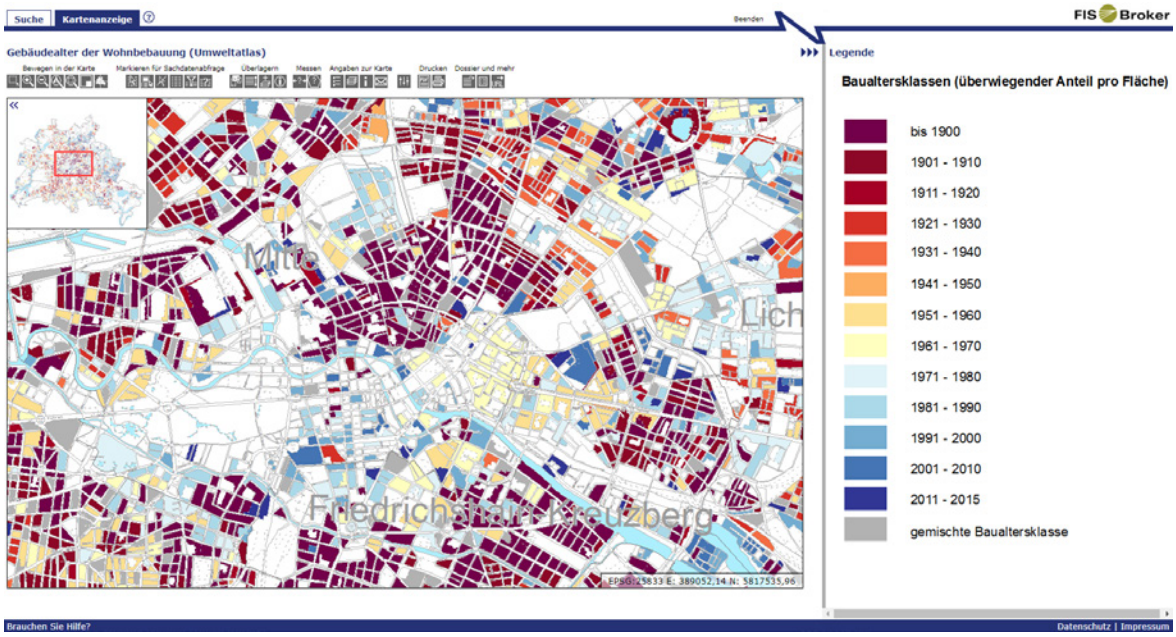
Figure 3: Unemployment rate for German cities



Source: German Urban Audit

The final action undertaken for this project concerned web applications and web services. More precisely, this concerned developing provisions for Urban Audit data so that they may be used to create tables and thematic maps (interactively generated), while providing opportunities to download data through a web-based information portal. As with some of the previous actions, the work was conducted by performing an analysis of the strengths and weaknesses of selected public information portals. Thereafter, a new information portal for the German Urban Audit was developed (see Figure 3) and information was shared so that individual cities could develop their own city portals: the first two prototypes were developed for Berlin (https://fbinter.stadt-berlin.de/fb/index.jsp; see Figure 4) and Freiburg im Breisgau (https://fritz.freiburg.de/Informationsportal/).

Figure 4: Construction date for dwellings in Berlin



Source: Geoportal Berlin

RESULTS

In Germany, cities are an important player within the statistical system: however, federalism and subsidiarity both establish obligations and opportunities to develop competences at a local level. The Merging Project provided a basis for taking further action when developing European infrastructures for geo-referenced data, for example within the realms of geo-coding addresses, harmonising spatial units, or implementing information systems, interfaces and information portals.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!mH87jx
Presentation: https://circabc.europa.eu/sd/a/9b8d92f2-70f8-4c3b-b8bd-7d5cdaf8d48f/08143.2013.004-2013.440_PresentationGISCO.pptx
DESTATIS — Federal Statistical Office: https://www.destatis.de/EN/Homepage.html
Mapping portal: https://web2.mannheim.de/urbanaudit/structuraldataatlas/index.html

Croatia

Croatian Bureau of Statistics

CROATIAN BUREAU OF STATISTICS

Merging statistics and geospatial information in Member States; 2013 project; final report 9 December 2015

KEYWORDS: census, visualisation, population, grid

PROBLEM

Prior to the initiation of this project there was a lack of know-how and experience for merging statistics and geospatial information within the Croatian Bureau of Statistics (CBS).

OBJECTIVES

This project provided the opportunity to merge 2011 census data about the population, households and dwellings with spatial information using geographic information system (GIS) tools. By doing so, one of the main objectives of the project was to integrate geographical and statistical information to increase the usefulness of data through grid-based population statistics, which would allow key policy initiatives (such as the Europe 2020 strategy or GDP and beyond initiative) to be defined better, measured more accurately, analysed and monitored at a local level.

METHOD

To provide users of official statistics with geographically referenced statistics, this project sought to make use of population grids as a tool to provide local information on societal developments and relationships. This type of data was considered to be of particular use for analysing phenomena which are independent of administrative boundaries, such as commuting or urban sprawl.

To deliver the project a working group was established and training was provided for several members of CBS staff, including:

- a workshop that provided theoretical and practical information pertaining to the production of grid-based statistics and practical exercises for using ArcMap10 software;
- a workshop on mapping statistics, that covered, among other things, theoretical and practical basics of mapping, information on effective dissemination, fundamental concepts of GIS and practical exercises;
- a workshop on analysing geospatial relationships.

When setting-up the system for grid-based statistics, the first step was to analyse the quality of data and the methodologies used elsewhere in the EU Member States. The population grid was established on the basis of geo-referenced national micro data and a country clip Lambert-Azimuthal Equal Area (LAEA) grid. The CBS generated grid cell identification codes for national micro data using GIS software, joining grid cell centroids with the country clip using ArcMap10 (the central application of ArcGIS software). The results were subsequently re-projected so that it was possible to populate the European ETRS89/LAEA grid and derive an ArcGIS shapefile with house numbers.

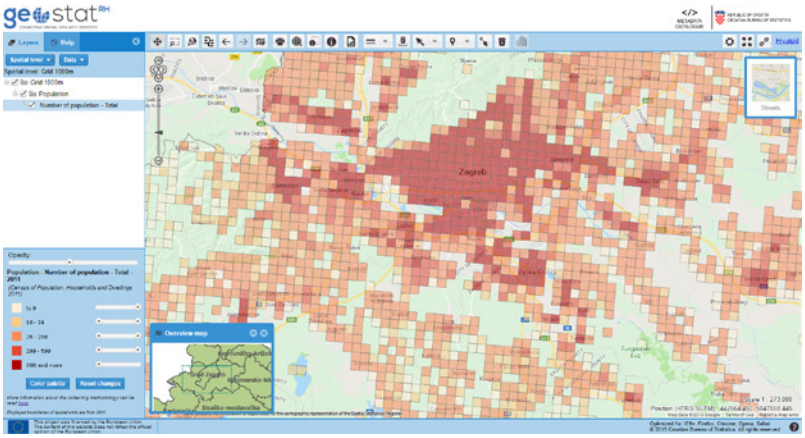
Once the CBS had a 1 km^2 grid set-up, use was made of algorithms to enable data protection and aggregation methods for point-based data or enumeration areas, generating a set of geospatial statistics within ArcMap10. The data produced were largely derived from CBS data sources, but also made use of information provided by the Croatian State Geodetic Administration (SGA). The data set was composed of both simple and complex statistics, together with accompanying analysis and included data about:

- population by sex;
- population by age (0-14, 15-64, ≥65 years);
- population by sex and age;
- population by activity status (economically active or inactive);
- population (aged ≥15 years) by highest level of educational attainment.

The final step in the project was to develop an application, the GeoSTAT portal, to assist in creating and publishing interactive maps for displaying grid-based results. Secondary data protection was implemented such that any grid cell with fewer than four persons had its information hidden for a set of main variables, while a threshold of fewer than 10 persons was applied to the indicators covering the population by activity status and the population by highest level of educational attainment.

The CBS produced a quality report on its grid-based statistics and a set of methodological guidelines for the Merging of statistical data of the Census of Population, Households and Dwellings 2011 and geospatial information in the Republic of Croatia was published on the CBS website at: http://www.dzs.hr/Hrv/publication/metodologije/metod_74.pdf.

Figure 1: Total number of inhabitants, by 1 km² grid

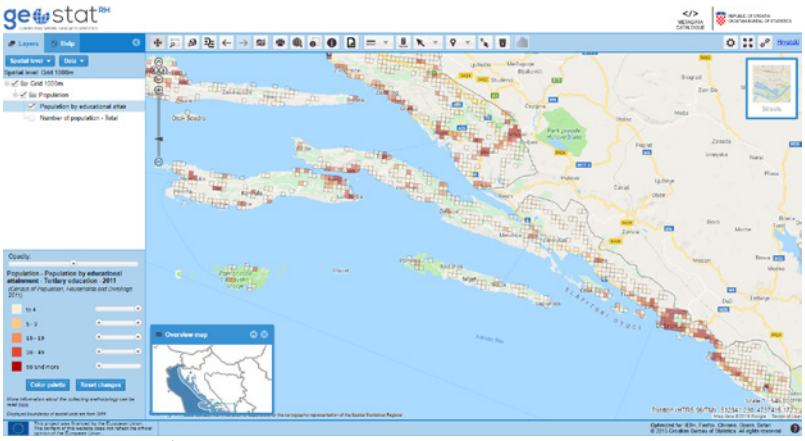


Source: GeoSTAT portal

RESULTS

The project was designed and documented such that the results could potentially be reused by other statistical entities within the EU Member States. Furthermore, much of the information and experience gained during the course of the project was used subsequently to develop a geoportal for the capital city, Zagreb, with information on its local infrastructure and spatial data (see: https://geoportal.zagreb.hr/).

Figure 2: Number of people with a tertiary level of educational attainment, by 1 km² grid



Source: GeoSTAT portal

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Gt49cu
Presentation: https://circabc.europa.eu/sd/a/ce320475-b2bc-48b3-a2aa-cd03ec809c34/08143.2013.004-2013.442_PresentationINSPIRE.ppt
Croatian Bureau of Statistics: https://www.dzs.hr/default_e.htm
Mapping portal: https://geostat.dzs.hr/

Italy

National Institute of Statistics

Standardisation and geo-coding of place names in the database of migratory flows; 2013 project; final report 15 September 2016

KEYWORDS: migration, residence permits, demographics

PROBLEM

A lack of information meant that it was not possible to visualise or analyse migratory flows into Italy from specific provinces of non-member countries (countries outside the EU) or to analyse the various areas in Italy where migrants from different countries were most inclined to settle.

OBJECTIVES

The main objectives of this project were to put in place systems to identify, map and analyse migrant flows into Italy of people from non-member countries. Using data collected from new applications for residence permits, the Italian Ministry of Interior managed a database covering the period 2012-2015 which contained information on each new migrant's place of birth.

In order to make efficient use of this information, the project identified four main activities, to:

- normalise the place names used for the birthplaces of migrants originating from the 20 most common partner countries;
- geo-code information pertaining to each migrant's place of birth;
- produce a set of maps detailing the birthplaces of migrants and their place of residence in Italy;
- organise a workshop so that the results could be presented, explaining how the quality of data might be improved and how tools developed during the project might be reused either by other countries wishing to conduct similar analyses or by different actors within Italy (for example, other ministries or statistical bodies).

METHOD

The starting point for the project was the database on new residence permits, which provided a source of information for detailing the birthplace of migrants arriving in Italy in terms of the country from which they came and the name of the city/town/village where they had been born. A decision was taken to perform an initial analysis for the 20 non-member countries that had the highest number of citizens arriving in Italy during the period 2012-2015. The top five partner countries during this period were Morocco, China, Albania, India and Bangladesh.

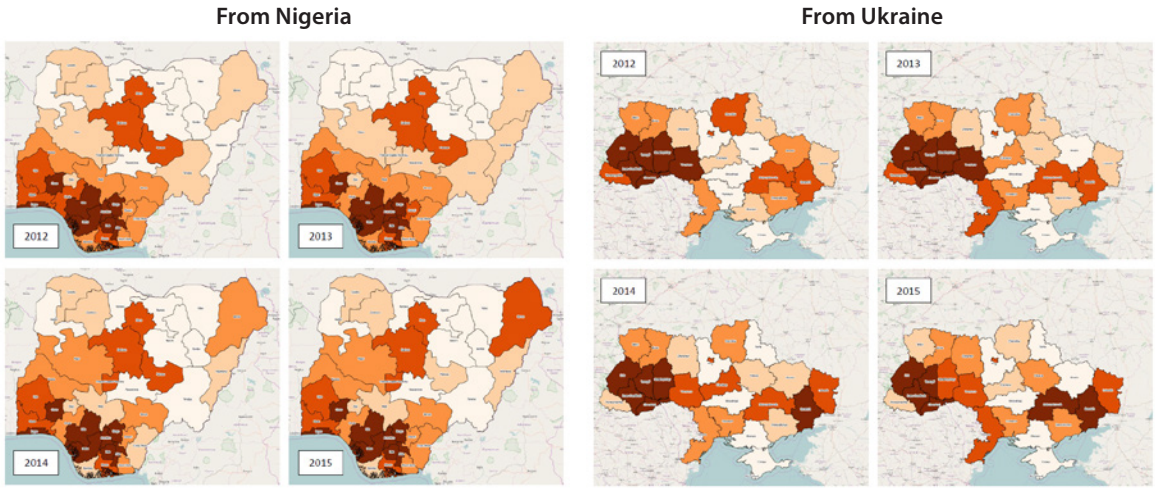
A normalisation process was designed so that the same birthplace names could be identified and used for each of the reference years. The first step was to create a unique table covering all four reference periods. Thereafter, superfluous or sensitive information was removed from the table to reduce its size and to make data processing more efficient, leaving a table structure that was composed of: migrant_ID, sex, country code, country name, birthplace, reference year.

Thereafter, the data was geo-coded — this operation was carried out directly within OpenRefine (a desktop application used for cleaning and transforming data sets). For the vast majority of the 20 partner countries there was a relatively small loss of information when normalising and geo-coding data entries: for example, the success rate for coding records for migrants originating from Morocco was 95.7 % and success rates of more than 90 % were recorded for most of the other partners. In a small number of cases records were lost as a result of the birthplace being such a small place that it was not contained in the Geonames database. Furthermore, there was a relatively low rate of success for normalising and geo-coding information on migrants originating from Ukraine, Moldova and Russia, where close to half of all records were discarded (this was often due to Cyrillic characters being employed or information on the birthplace being replaced by the country of birth).

RESULTS

Having processed the data, the next step was to produce maps for each of the 20 partners and the four reference years. The maps detailed the flow of migrants to Italy from local administrative areas. In some cases there was little fluctuation in the flows between different years, although there were some specific cases where large fluctuations could be seen — often resulting from socioeconomic and/or political crises — for example, there was a considerable increase in the number of migrants moving from north-east Nigeria to Italy in 2014-2015 (which could be linked to increased levels of unrest in the area associated with the terrorist activities), while there was also a sizeable increase in the flow of migrants from eastern provinces of Ukraine (following the conflict with Russia).

Figure 1: Migratory flows to Italy, 2012-2015

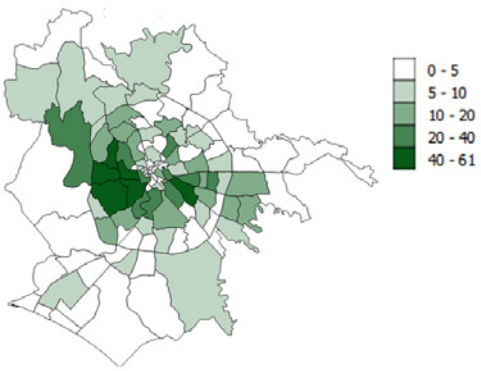


Source: ISTAT

Alongside maps representing the origins of migrants who made their way to Italy, the data set was also used to construct information concerning geographic poles of migrant residence, highlighting specific (neighbourhoods of large) cities where migrants from a particular country tended to settle.

The final task carried out under this project was the organisation of a workshop in June 2016 to present the main results and to provide an opportunity for sharing knowledge, ideas and new perspectives that might be adopted when analysing migration flows.

Figure 2: Flows of migrants from India settling in Rome, 2012



Source: ISTAT

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!jy94nG
Presentation: https://circabc.europa.eu/sd/a/76b4d435-6a21-425a-8383-1c498dd4c4bc/08143.2013.004-2013.443_PresentationGISCO.pptx
Istituto Nazionale di Statistica (ISTAT): https://www.istat.it/en/
Studying migration routes, new data and tools — workshop documents: https://www.istat.it/en/archive/186853

Austria

Statistics Austria

Census 2011 — enriching commuter statistics, 2013 project; final report February 2016

KEYWORDS: commuting, local administrative units (LAU), register

PROBLEM

From 2011, variables on individual commuting patterns — such as commuting time or distance travelled — ceased to be available within the Austrian statistical system as Statistics Austria moved to a register-based census. Initial attempts to estimate variables for measuring commuting had implausible results, leading to a search for new methods.

OBJECTIVES

The goal of the project was to develop a set of methods to improve the quality of estimates for various measures of commuting. The population census provided a record — for each person — detailing where they lived (through a building_ID for their home address) and where they worked or went to school (through a work_ID or school_ID), supplemented by information for each individual's economic activity status and other demographic data.

METHOD

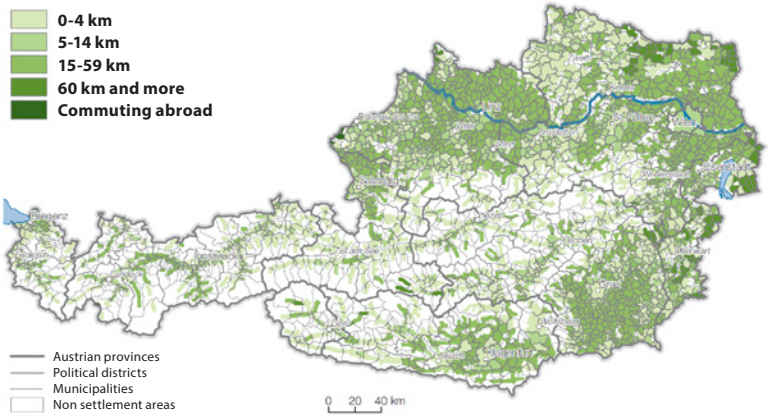
The project was based on using ArcGIS10.1 software with a network analyst extension. TomTom was chosen as the data source for information on streets and routing with an in-house solution to detail numerous variables on each street (such as one-way streets, types of road, speed limits, or whether the road was in the countryside or a built-up area). The model builder and scripts were developed in Python 2.7 with SQL-scripts for tabular results.

Commuters were defined, for the purpose of this project, as employed or self-employed persons as well as pupils or students, who travelled between their place of residence and their place of work or education. As such, unemployed and other economically inactive people, as well as those who worked from home were excluded. Furthermore, people living in Austria and working abroad or people living abroad and working in Austria — frontier workers — were also removed from the population under consideration (as information on their foreign address was not always available).

The buildings and dwellings register (BDR) run by Statistik Austria contained information on the details of land, buildings and dwellings as well as structural data, such as the x,y coordinates of each building. It was combined with other data sources, such as the central register of residents, tax registers or education registers to encode personal_IDs and building_IDs that were geo-referenced.

To make use of the data from various registers, these tables had to be joined through the use of scripts so that commuting journeys made by approximately 4.3 million out of 8.4 million Austrian residents could be analysed in more detail. The coordinates for building_IDs were extracted and re-projected to allow them to be used with other data sources and the street network system, with the main goal being to model distances travelled and commuting times for each commuter.

Figure 1: Most frequent distance travelled by economically active commuters, 2013



Note: Vienna is presented as one municipality.

Source: STATISTIK AUSTRIA, Register-based Labour Market statistics 2013 by municipality as of 2013

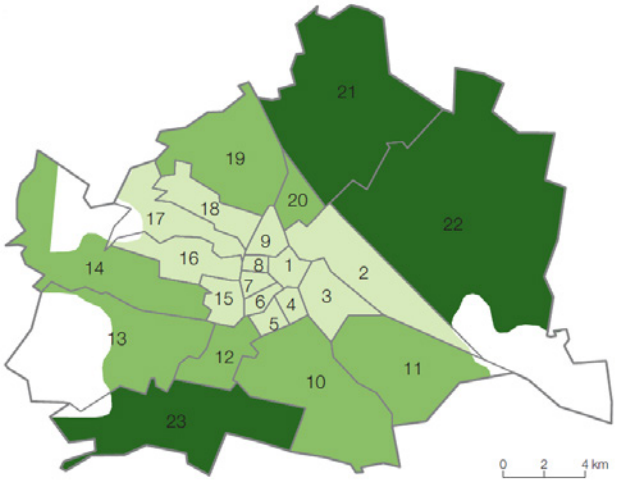
There were several issues encountered when trying to perform this task:

- some pairs of destinations had zero distance when commuters were found to live and work on the same stretch of road;
- some buildings had the wrong coordinates (for example, the coordinates of a primary school in Vienna were found to be more than 15 km away from their true location);
- some minor alpine, forestry or private roads were found to be missing from the road network;
- some buildings were found to be located up to 5 km from the nearest road;
- transit routes through neighbouring countries were ignored in favour of the national road network.

Otherwise, it was a relatively easy task to calculate the shortest route for each commuting pair, whereas it was more difficult to model the optimised route and the driving time required. This difficulty resulted from a lack of information on speed profiles for various stretches of road from TomTom, so a model had to be developed to provide a realistic estimate of driving times (this was based on road conditions that were neither congested nor totally clear). To improve the model, a fictive set of routes was calculated using various routing engines and the results from these were used as the basis for further calibration, care being taken to select a representative sample of fictive routes that covered routes to/from and across built-up areas, suburbs and the countryside.

With the data, the street network and a speed model prepared, the next challenge was to find the fastest routes for the 4.27 million commuters, a task that required considerable computing power and a lengthy period of time for the scripts to be run. Once the results had been obtained, attention turned to how best to disseminate the information and aggregate house-to-house information to identify commuting zone matrices, identifying in-commuting zones (clusters of high workplace density) and out-commuting zones (residential, suburban areas).

Figure 2: Most frequent distance travelled by economically active commuters, municipal districts of Vienna, 2013



Source: Labour market register, Statistik Austria

RESULTS

The results obtained for commuting distances and commuting times were tested and found to work well. On the other hand, identifying the optimal level of detail for analysing commuting matrices was less clear.

The work carried out as part of this project, including the algorithms created, may be reused in years to come or alternatively by statistical entities in other EU Member States, as Statistik Austria documented both the routing model and the scripts.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!MK67rd
Statistik Austria: http://statistik.at/web_en/statistics/index.html

Slovenia

Republic of Slovenia
Statistical Office

REPUBLIC OF SLOVENIA
STATISTICAL OFFICE

Merging statistics and geospatial information in Member States, 2013 project; final report 2 March 2016

KEYWORDS: mobile communications, big data, grid, legal

PROBLEM

Although the Statistical Office of the Republic of Slovenia (SURS) had considerable experience in using and disseminating various register-based geo-referenced statistics, this expertise did not extend to temporal analyses.

OBJECTIVES

This project looked at a specific challenge for official statistics, namely, to make use of already existing data managed by mobile network operators to open-up new fields of analyses. The main objectives of the project were to:

- examine the legal framework necessary to permit optimal/secure acquiring, handling and dissemination of data from mobile network operators;
- improve the integration of geo-information and geo-referencing into the statistical production process;
- illustrate how links between geo-information and statistical information could be used to potentially provide additional value and new information.

METHOD

SURS obtained data on mobile telecommunications from the second largest mobile network operator in Slovenia. The data set covered the period between 1 May and 31 October 2014 and contained information on all events that were initiated by mobile users, as recorded through geo-referenced x,y coordinates of mobile base stations (antennae). The project explored how such large volumes of geo-referenced information might be handled using standard geographic information system (GIS) processes to open-up new forms of temporal analyses for population mobility (as measured by changes in the location of mobile phone users).

The data set that was obtained for the project included the following information:

- the x,y coordinates of base stations used to send/receive radio signals to each mobile phone user;
- a unique user_ID that was anonymised;
- the exact time of initiated events (for example, when making a call, or sending an SMS).

The distribution of mobile base stations tends to be denser in urban and commercial areas or along major transport arteries and consequently scarcer in more rural areas. This issue might, at least to some degree, be rectified if the project were to be extended so it covered more than one mobile network operator.

Through the processing of mobile data, SURS acquired experience in various activities — from data transmission, secure and auditable processing, through to the development of GIS applications and data visualisations. The data were stored in tables (one for each month) with an auditing system centred on an ORACLE database containing in excess of a billion records. The main challenge was in relation to the storage of data (an issue that may prove to be of even greater relevance if the project is, one day, extended to cover the regular transmission of data from all mobile operators in Slovenia).

As SURS wished to compare the data on the position of mobile phones with register-based data, it was necessary to find a common territorial division whereby both data sources could be joined. To do this, the areas between individual base stations were analysed to determine a set of Voronoi polygons providing an estimation of the boundaries in coverage between base stations.

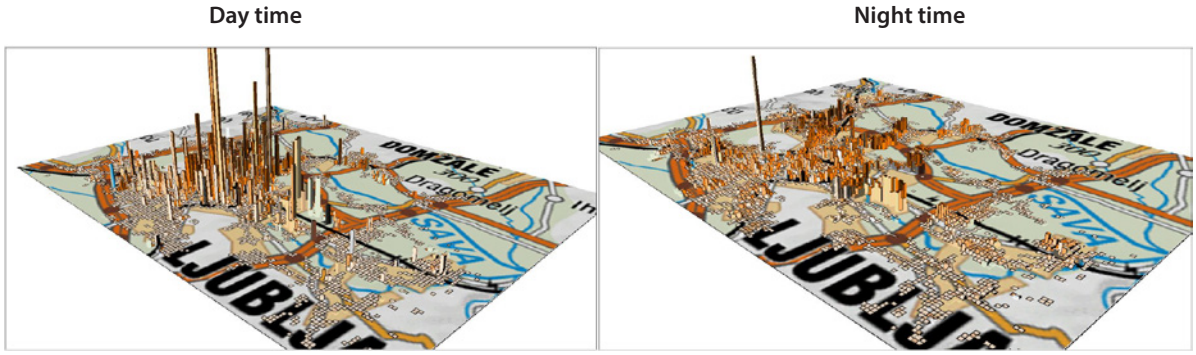
The Slovenian information commissioner dissuaded SURS from directly linking data on an individual's location/the position of their mobile phone with register-based data, so as to avoid any potential disclosure of user identity. Therefore, SURS developed a model based on a 500 m^2 grid, estimating temporal population densities for each cell. Point-based locations of the base stations were merged with information for municipalities to generate new municipality-like divisions based on the Voronoi polygons around each base station. This solution allowed SURS to develop analyses combining register-based data (such as that for employees or actual residences) with mobile phone data detailing an individual's position during the day. The temporal dimension of the data set placed these statistics in a new perspective as the data on mobile phone locations were captured in real-time, thereby highlighting temporal patterns which could be used for a variety of applications, for example, an analysis of the average time taken to get to work or preferred routes taken by commuters.

RESULTS

On average 9.9 % of mobile phone users covered by the project generated at least three quarters of their events through a single base station and 41.1% generated at least half of their events through a single base station. Aggregating this information to the level of municipalities, almost half (47.0 %) of all users generated at least three quarters of their events in a single municipality, while 82.6 % of users generated at least half of their events in a single municipality.

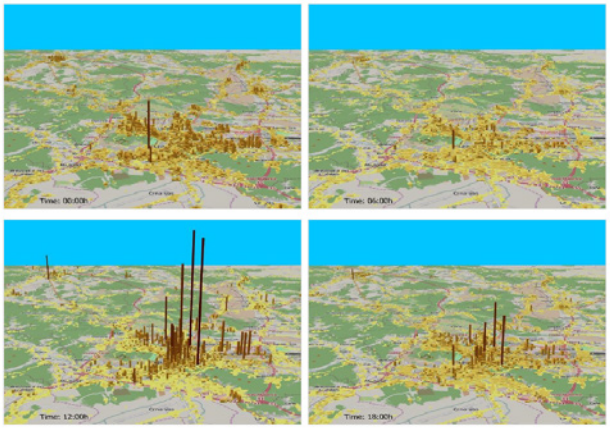
The results also showed that there was a noticeable influx of commuters into some of Slovenia's major cities each morning, which led to a considerable change in their populations. Figure 1 contrasts day-time and night-time population densities for the territorial boundaries of Ljubljana (the capital of Slovenia) during the period 1 May–31 October 2014; the day-time population having been measured at lunchtime (12:00h-13:00h), while the night-time population was measured after most people had gone to bed (00:00h-01:00h). The information presented is based on administrative data (for the day-time population) and a central population register (for the night-time population) and was calibrated using hourly patterns observed in GPS tracking from mobile phone data. The general pattern of considerably more people being in the city centre during the day-time is clearly apparent when contrasting the two maps, whereas the night-time population grew in some sub-city districts away from the centre, as well as surrounding suburban areas.

Figure 1: Day and night time populations, 500 m^2 grid, Ljubljana, 2015



Source: SURS

Figure 2: Estimated hourly distribution of the population, Ljubljana, 2015



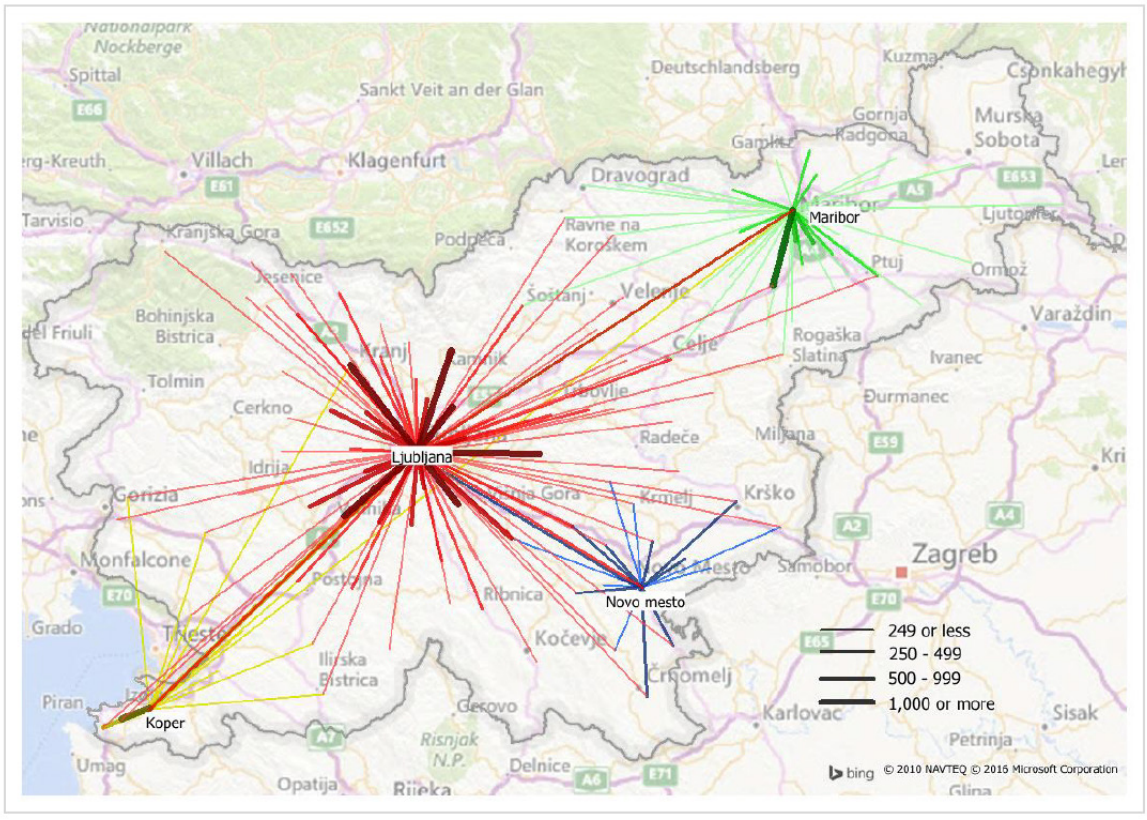
Source: SURS

The results of tracking mobile phone users were also used to estimate the population of Ljubljana, at different points of time during the day (each hour). Figure 2 shows information for four different points in time (00:00, 06:00, 12:00 and 18:00), highlighting how the population of the central business district expands during the working day.

Information on the mobility of mobile phone users was also used to estimate the attractiveness of various municipalities/cities. Figure 3 presents information on the attractiveness of Ljubljana, Maribor, Novo mesto and Koper, based on the number of commuter inflows into these cities during the working week (Monday-Friday).

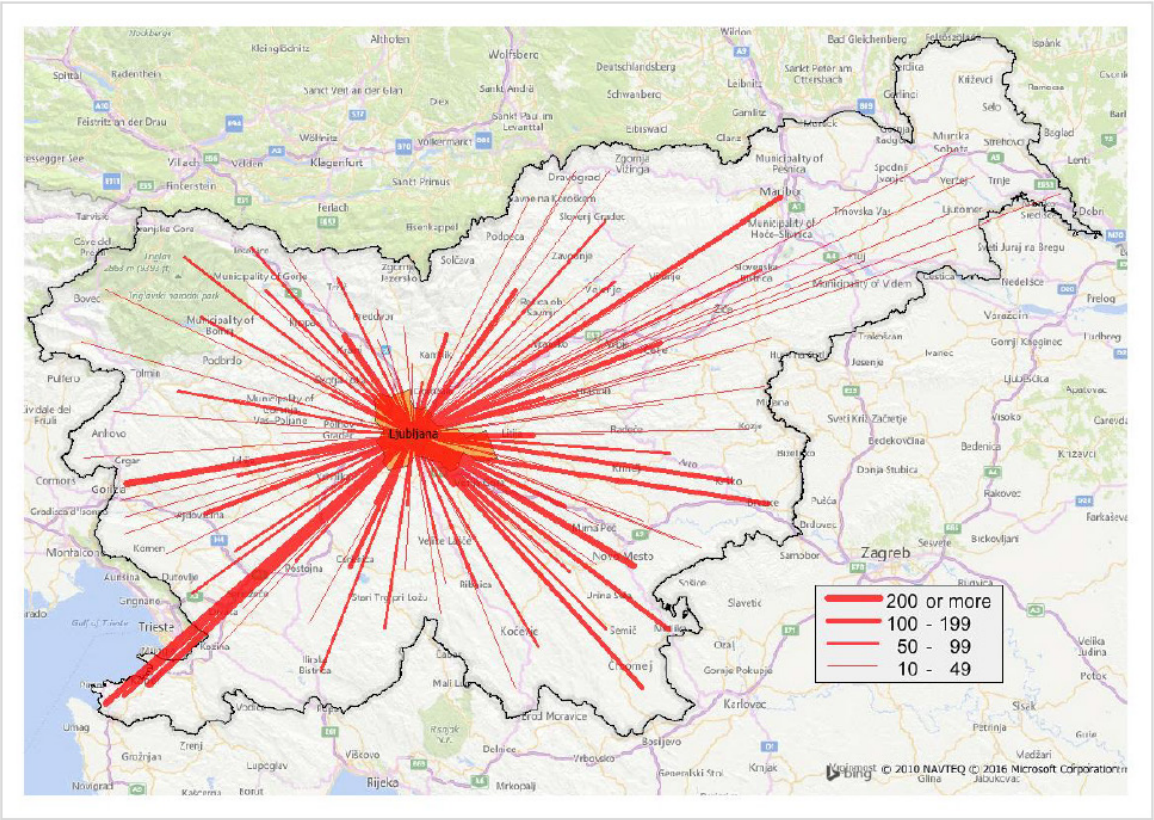
A final example demonstrates how information for people usually resident in Ljubljana may be analysed to ascertain if they leave the city at weekends (perhaps to visit family or to make use of a secondary dwelling in the countryside or on the coast). Figure 4 shows information on weekend locations that are favoured by Ljubljana residents.

Figure 3: Attractiveness in terms of daily commuter inflows, 2015



Source: SURS

Figure 4: Usual location of Ljubljana residents on Saturdays and Sundays, 2015



Source: SURS

Given that data access was a considerable issue when setting-up the project — with lengthy discussions over the potential use that might be made of this big data source — SURS devoted a considerable amount of time following the study to promote the activities that were undertaken, with the goal of expanding the project to a range of new applications including, for example, studies of commuter, tourism or transport patterns, or applications relating to civil protection and/or disaster relief.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!YB73tw
Presentation: https://circabc.europa.eu/sd/a/7735841a-ab1c-418f-9239-6c6ddebb998b/08143.2013.004-2013.444_PresentationGISCO.pptx
Republic of Slovenia Statistical Office: https://www.stat.si/StatWeb/en
Slovenian statistical regions and municipalities in numbers: https://www.stat.si/obcine/en

Finland

Statistics Finland

Statistics Finland

Spatial statistics on the web, 2013 project; final report 29 January 2016

KEYWORDS: geoportal, visualisation, small area statistics

PROBLEM

Traditionally, the compilation of small area statistics was an activity for which Statistics Finland charged a fee. New standards and technological developments led to a decision being taken to explore the possibility of providing these statistics for spatial analysis as free, open data.

OBJECTIVES

This project was carried out in cooperation between Statistics Finland and the National Land Survey of Finland during 2014 and 2015. Its principal aim was to implement an open source web application enabling the spatial analysis of data through an easy-to-use and innovative web application that was based on the use of spatial data infrastructure via INSPIRE-compatible web feature services (WFS) and web map services (WMS) providing direct access for reading, writing and updating data and for the transformation of data into maps.

METHOD

A web application was built using Oskari (open source software with a dual European Union Public Licence (EUPL) and Massachusetts Institute of Technology (MIT) license). It was developed to make using spatial data infrastructures easier, through providing tools and functionalities for distributing geostatistical information and analyses, such as allowing users to define their own maps and to embed and distribute these across their own websites.

Paikkatietoikkuna is a geoportal (available at https://kartta.paikkatietoikkuna.fi/?lang=en): it is an application developed through case story examples that were given to the project development team to stimulate improvements. One example of such a case story was:

"As a user I want to find out how many people live in the vicinity of a point feature. I want to be able to select the point by clicking on the map or by giving an address. I want to create a buffer around the chosen point and aggregate the total amount of people living inside the grid cells falling inside the created buffer area. Later I want to save and publish the results on my own map."

As a result, the goals and aims of the project included, to:

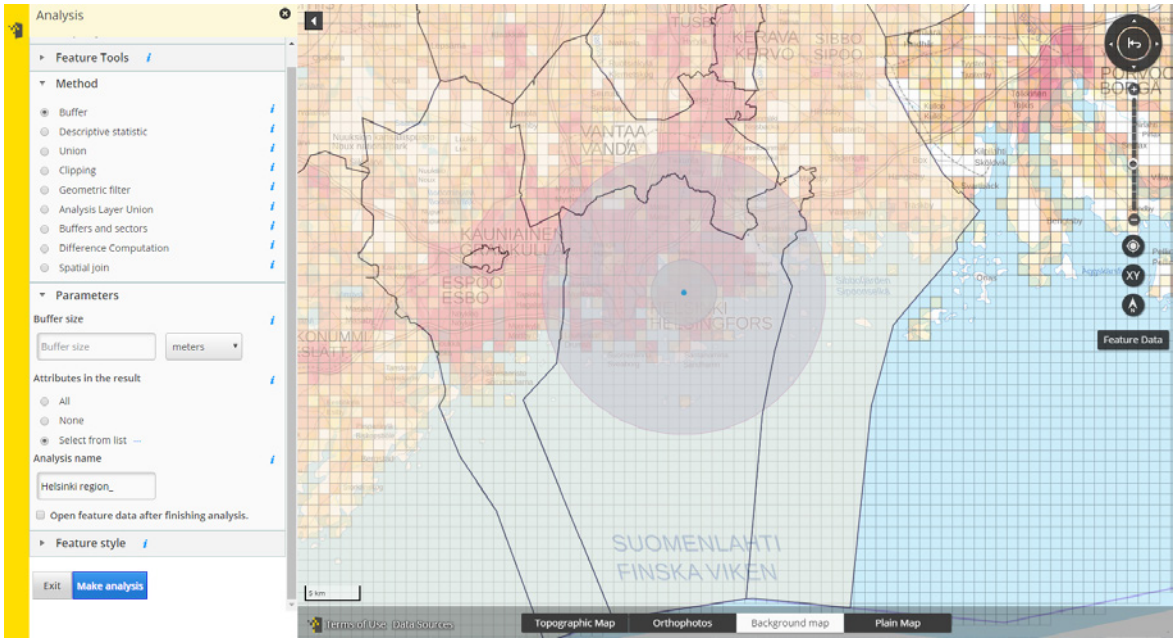
- increase the usability of grid-based statistics (and later small area statistics);
- illustrate how the integration of geospatial and statistical information may be used to provide added value by creating new information;
- offer spatial analysis tools within a web application.

At the start of the project a decision was taken to base development work on employing INSPIRE data specifications, metadata, service sharing and geo-server technologies. As such, data were collected only once and kept in a single location where they could be maintained most effectively. The data was structured and documented so that it could be combined seamlessly with other spatial information from different sources and shared with various users and across applications. The data was stored so that information collected at one level could be shared, analysed and published at other (more highly aggregated) levels. A set of information was made readily available, allowing users to easily find what geographic information was available, how it might meet their particular needs and under which conditions it could be used.

The data sets collated as part of this project included:

- municipality-based statistical units;
- population by municipality-based units;
- educational institutions;
- industrial facilities;
- 1 km^2 grid statistics.

Figure 1: Screenshot of Paikkatietoikkuna application with an example of a point feature and buffer



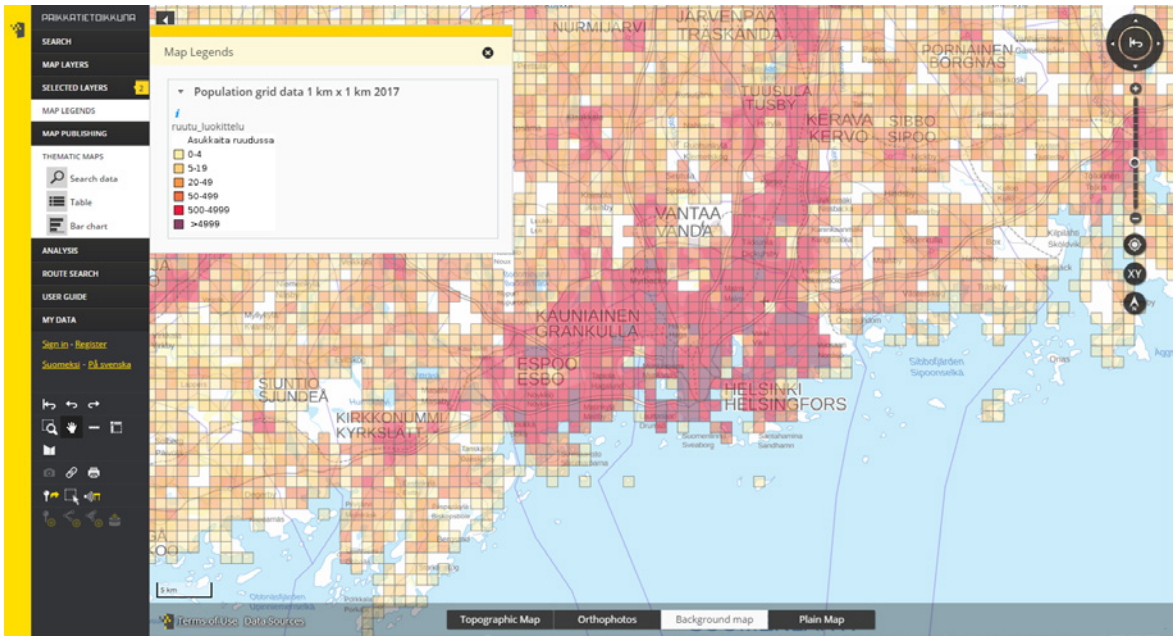
Source: Paikkatietoikkuna, Statistics Finland

RESULTS

Upon completion of the development work, the application was able to provide functionalities for analysis, user-defined buffer sizes, descriptive statistics, unions (the joining together of selected features into one unique feature), clipping (user-defined clipping layers to restrict/filter results), geometric filters (allowing partial or total intersections of layers), difference analysis between unique map layers (for example, subtracting the population of one period from the population of another), or heat maps.

Users may log-in to the application and build customised maps from various data layers. If the resulting map is saved it can then be embedded into a webpage by means of an http link created for each user-defined map. Such maps are dynamic insofar as anyone who visits the link can move the focus of the map, zoom the map, or click on various map features to obtain more information, or edit the information displayed so that it caters for their own analytical needs.

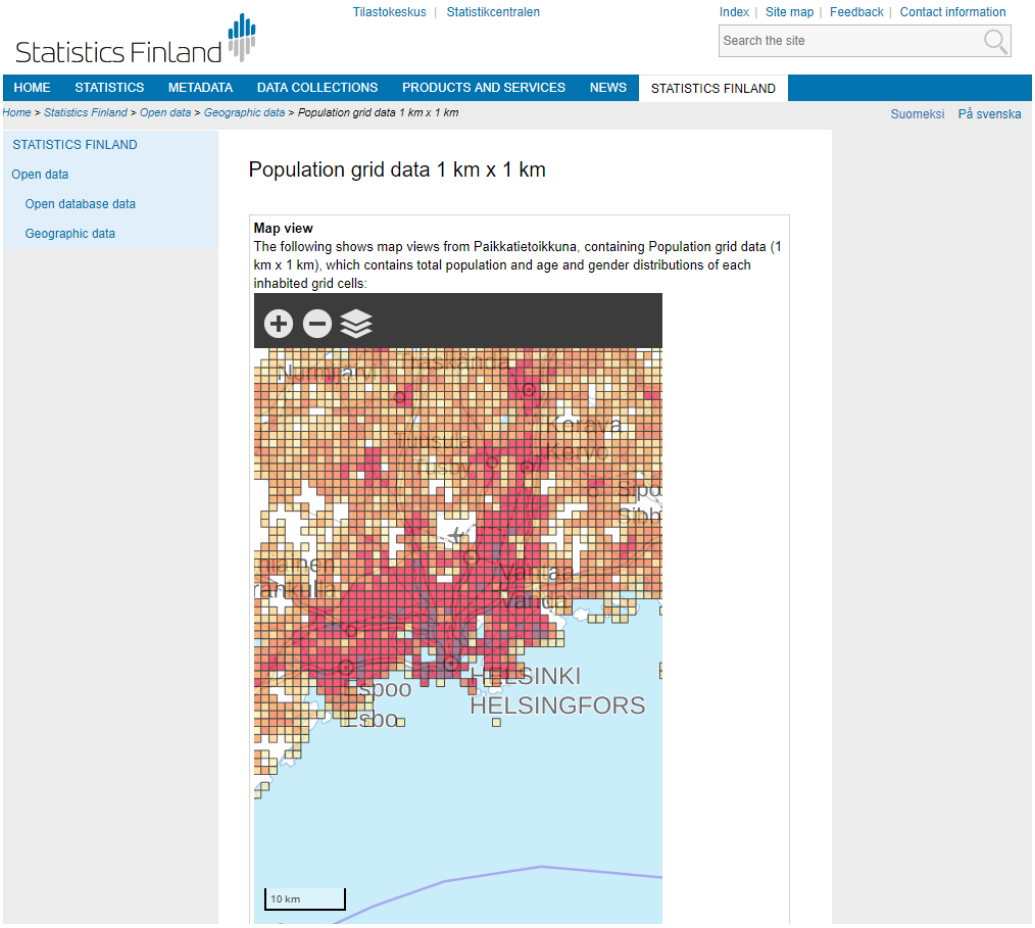
Figure 2: Population, 1 km² grid data, 2017



Source: Paikkatietoikkuna, Statistics Finland

An example is provided as to how a dynamic, embedded map can be presented. It is based on information published on Statistics Finland's own website. The example shows an extract of the same data presented in Figure 2.

Figure 3: Embedded map on the Statistics Finland website, population, 1 km² grid data, 2017



Source: Paikkatietoikkuna, Statistics Finland

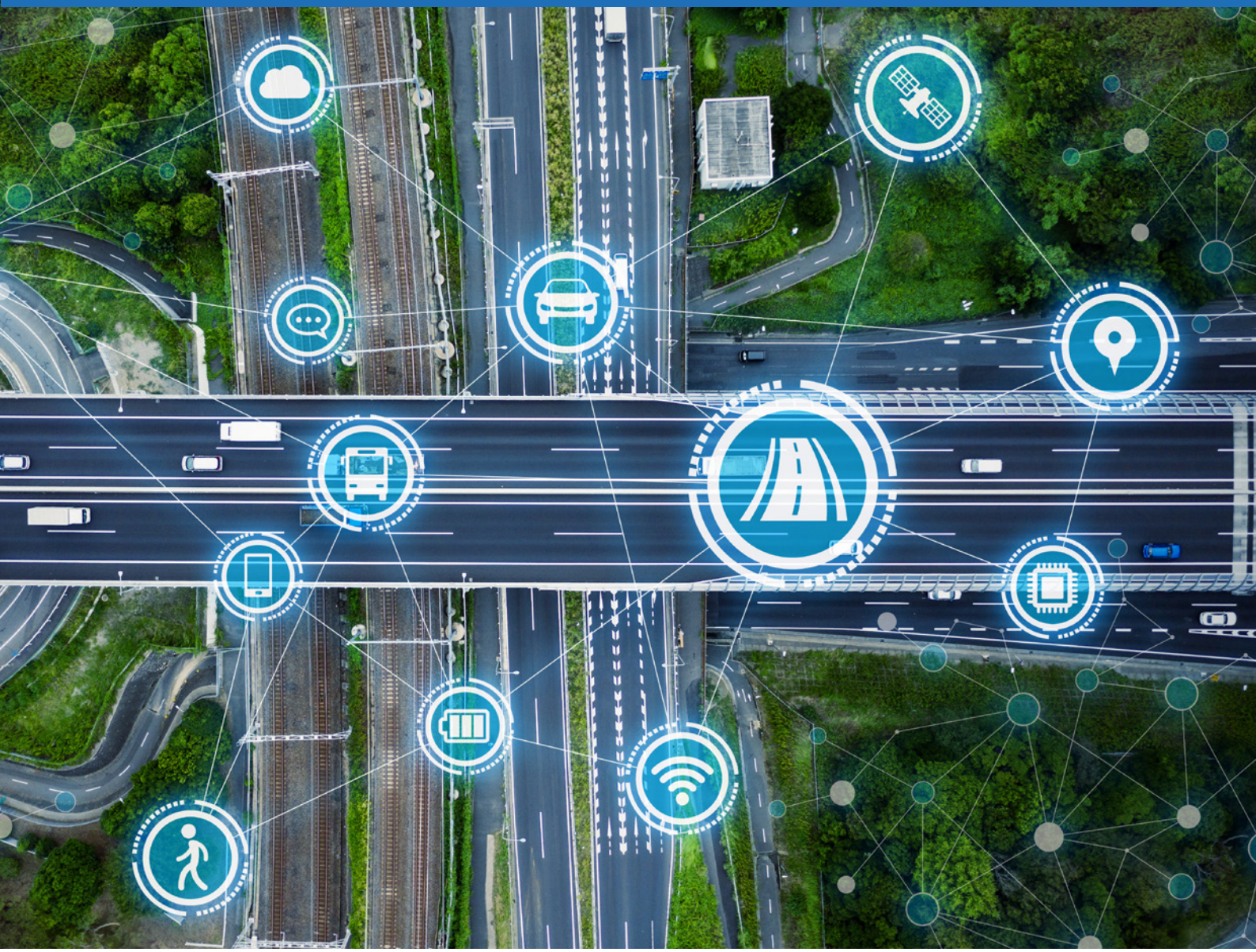
Through the development of Paikkatietoikkuna, an opportunity has been provided for users to visualise Statistics Finland's geospatial data. One downside, identified upon completion of the project, was the need to make the interface more user-friendly, as it was considered to cater principally for professional users who already had some prior experience of using spatial analysis tools.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!QK83py
Presentation: https://circabc.europa.eu/sd/a/280d03b9-e570-4b69-934a-9fec0fb30768/08143.2013.004-2013.441_PresentationINSPIRE.pptx
Statistics Finland: https://www.stat.fi/index_en.html
Paikkatietoikkuna: https://kartta.paikkatietoikkuna.fi/?lang=en
Population grid data: http://www.stat.fi/org/avoindata/paikkatietoaineistot/vaestoruutuaineisto_1km_en.html

C

2014 projects



Estonia

Statistics Estonia

Merging statistics and geospatial information in Member States — an address data system; 2014 project; final report December 2016

KEYWORDS: register, population, analytics, agriculture, grid

PROBLEM

There was no harmonised system to help generate a complete list of addresses for various statistical units. There was a lack of development for spatial analyses, for example, a methodology to disseminate statistics on economic and agricultural units through linking statistical information to geodata on economic units and agricultural holdings.

OBJECTIVES

Action 1: to improve the integration of spatial information and geo-referencing in the statistical production process (including survey design).

Action 2: to illustrate how linking geo- and statistical information and corresponding metadata may provide additional value and create new information.

METHOD

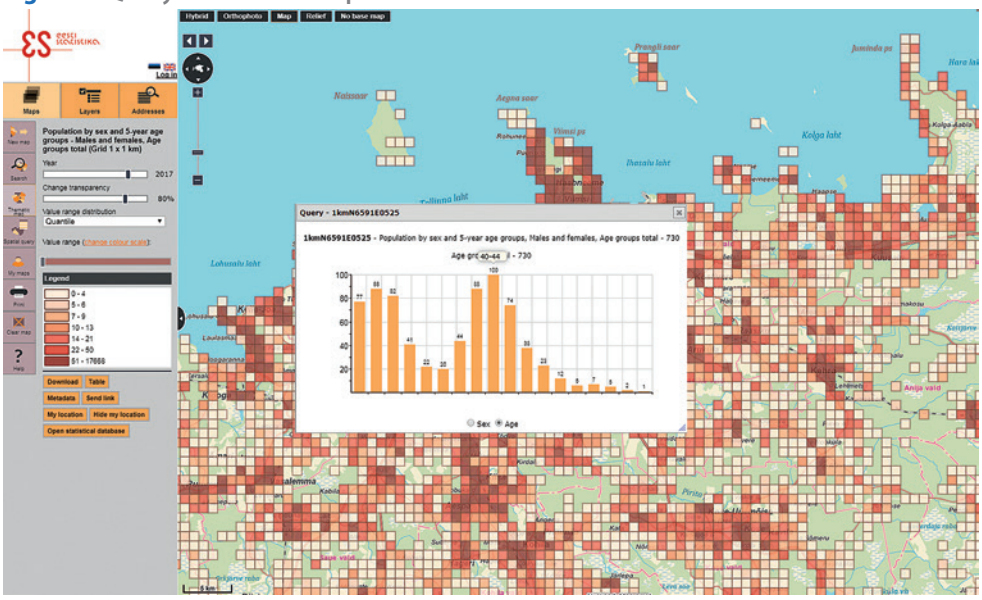
Action 1: during the project, data for statistical units in the statistical business and farm registers (SBR and SFR respectively) were geo-referenced. This initially involved: i) cleaning the addresses, for example to correct non-valid addresses or to add missing EHAK (the classification of Estonian administrative units and settlements) codes; ii) normalising the addresses to follow precise spelling and punctuation rules and to impose a specific structure made-up of eight components, and; iii) matching the addresses to an existing (but still under development) address data system (ADS). As such, a methodology for the treatment of non-valid addresses was developed and implemented. Cooperation with the Estonian Land Register (responsible for ADS) on the treatment of non-valid addresses in ADS continued.

A methodology was developed to create and update an integrated statistical population register for use as a frame for demographic statistics, household surveys and a register-based population census. The results of the 2011 population and housing census were adjusted for under-coverage using model-based estimates. The database was then updated annually based on registered changes, such as for births, deaths or migration (the latter covering all address changes). In 2016, the method of updating was changed to tackle problems of over-coverage resulting from incorrect migration information. The new system was based on integrating information from 14 administrative registers to look for so-called 'signs of life'. Each person was then assigned an index based on the number of registers which recorded signs of life and those with a low index value were assumed to no longer be residents.

A population database (as of 1 January 2016) was created including every person's place of residence with their full address linked to the land register (ADS). Where an address could not be linked to a place of residence it was linked to the centroid of a settlement. A 1 km^2 grid map of the population analysed by sex and by age group was produced and released through an application based on a statistical map.

Action 2: the geo-referencing of all data in the SBR and SFR made it possible to develop a system of spatial analyses and publications for data on the business population, business demography and the agricultural population. Note that data for business units were assigned to the economic unit's legal address — rather than local units — for reasons of data availability; this was explained in the accompanying metadata. Equally, farm land and buildings may be dispersed across multiple grid cells or administrative or statistical units, and therefore data for each farm was assigned to the central point of each holding. The data on economic units derived from the SBR were presented in maps, both online and in publications. The dissemination of geo-referenced data on agricultural was tested but not implemented; this included the production of maps based on a 1 km^2 grid showing the number of holdings in each grid cell as well as the average age of natural persons who were farm holders within each grid cell. Before publishing such information, some issues of confidentiality needed to be addressed.

Figure 1: Query of thematic map



Methods were developed in order to publish data for grid cells. Point-based statistical data were aggregated for grid cells based on linking coordinates with individual grid cells. These data were treated for confidentiality and metadata was also prepared.

RESULTS

The integration of spatial information and geo-referencing in the statistical production process was improved. Statistical business and farm register were fully linked to the ADS and through this the geo-referencing of data for statistical units in each of these registers was completed.

National population grid data were updated following the population and housing census for 2011 and a map of the population grid was published.

Adding value was demonstrated by illustrating the possibility of creating new information through linking geo- and statistical information and corresponding metadata. The basis for spatial analyses and a methodology for the dissemination of statistics on economic and agricultural units through a map application were developed.

New maps were displayed in the map application, based on 1 km^2 grid cells.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Hg68pf
Statistics Estonia: https://www.stat.ee/en
Statistical office's maps: https://www.stat.ee/maps
Mapping portal: https://estat.stat.ee/StatistikaKaart/VKR

France

National Institute of Statistics and Economic Studies

Institut national de la statistique et des études économiques

Insee

Mesurer pour comprendre

Towards a French address register; 2014 project; final report 24 February 2017

KEYWORDS: address, analytics

PROBLEM

In France, there was significant room to improve the management of addresses as geographical data items. The problem originated, at least in part, from the fact that municipalities (communes) are responsible for naming and numbering roads and streets at a local level (36 000 different municipalities), without any national operator to coordinate, centralise and standardise them.

OBJECTIVES

Action 1: to assess the strengths and weaknesses of various databases that were available for addresses.

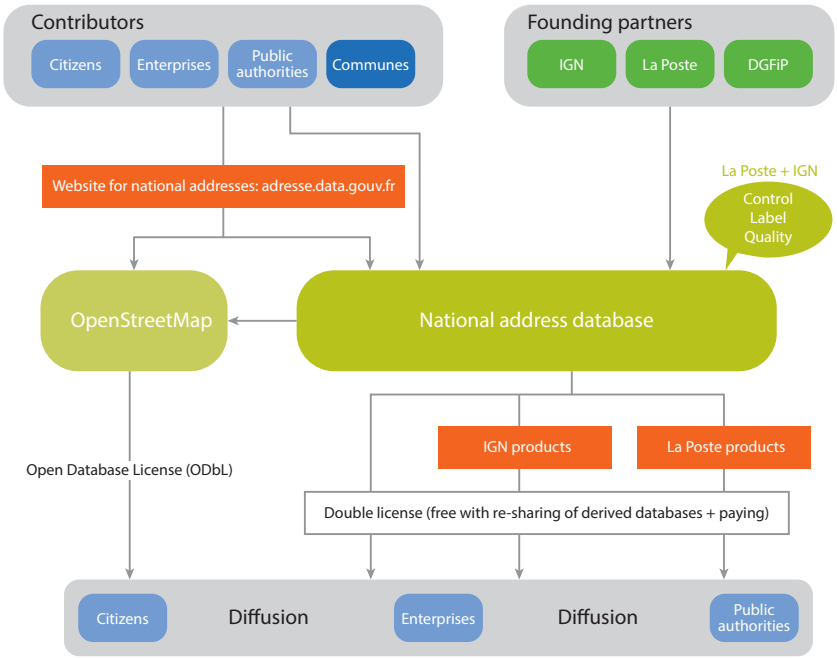
Action 2: to develop a prototype address register to improve and supplement the existing databases within INSEE.

Action 3: to evaluate the conditions for producing and updating — for statistical purposes — a common address register (RCA).

METHOD

Action 1: municipalities (communes) are responsible for naming and numbering roads and streets and there are approximately 36 000 of these in France, without any national address operator to coordinate, centralise and/or standardise addresses. In some (often rural) areas, the addressing system is considered to be under-developed and unsatisfactory.

Figure 1: Simplified diagram of the target organisational structure for the BAN



Many address databases exist, most notably those for: the state postal operator (La Poste); the French national mapping agency (Institut national de l'information géographique et forestière, IGN); the French tax authorities (the Direction Générale des Finances Publiques, DGFiP; which is also responsible for drawing up the cadastral map), and; the statistical office, INSEE (which has an address control list). Some local authorities also have databases for their own use covering local areas, either through their own portals or a shared public portals (based on the Etalab mission), while an association — OpenStreetMap (OSM) France — also developed a publicly available geocoded database.

Numerous non-geocoded databases, mainly of administrative origin, contain address information that may prove useful for the compilation of an address register. A major development within this context was the development of a national address database, set up by La Poste, IGN, the Etalab mission and OSM; this was launched in April 2015.

Action 2: three geocoded address databases were analysed — INSEE's address control list (ACL), the DGFiP's address database (derived from tax sources) and the IGN's address database.

There was no identifying feature that was common to these three databases that could be used to match them. Therefore, a textual comparison of the components of the addresses was required. INSEE had an in-house application that was used to perform this merging operation: it assigned a probability score for potential matches. The information was standardised: for example, the word 'boulevard' was searched for and also checked/recognised from its abbreviations, while insignificant words such as articles (for example, 'le' or 'la') were ignored. The comparison of addresses from three databases showed that:

- 76.9 % of ACL addresses were found without any ambiguity in both the IGN and DGFiP databases;
- 6.8 % of ACL addresses were found without any ambiguity in the DGFiP database only;
- 5.0 % of ACL addresses were found without any ambiguity in the IGN database only;
- 11.3 % of ACL addresses were not found in either the IGN or the DGFiP databases.

The failure to match ACL addresses with the IGN and DGFiP databases may be largely explained by difficulties in recognising the road name as part of an address (81 % of unmatched ACL addresses), while the remainder of the difficulties were due to house number correspondence problems. Problems matching road names were often related to different types of roads (for example, drive, street, avenue), parts of the name missing (sometimes related to the maximum number of characters, which varied between databases), treatment of dates or numbers in a road name (for example, with ordinal numbers presented differently), or the use of secondary names (when a building had more than one address).

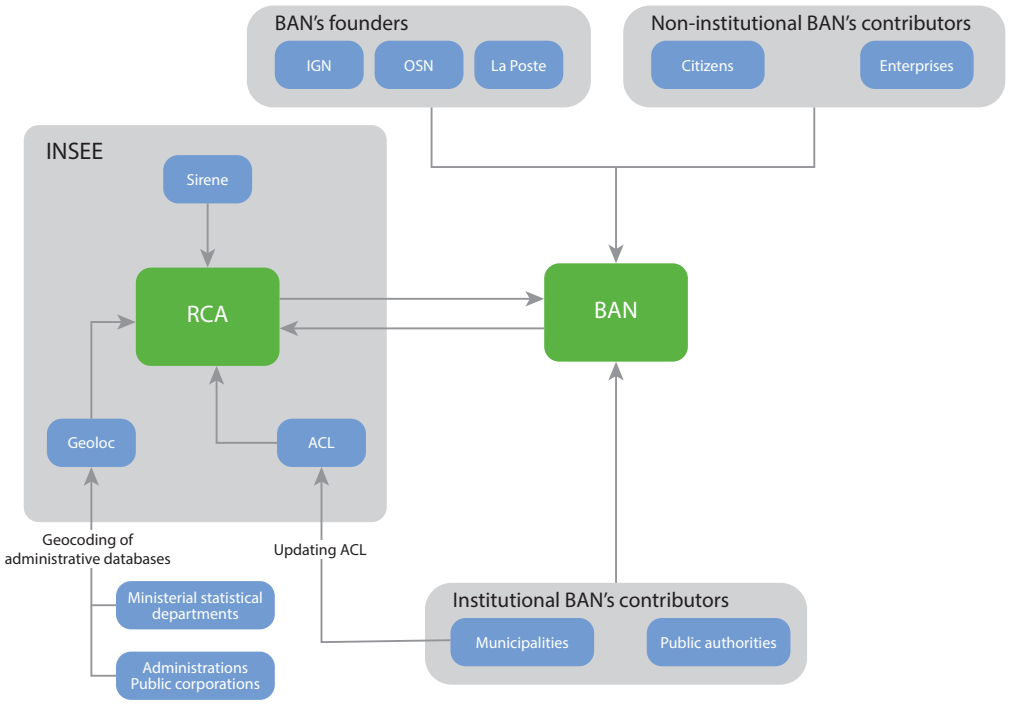
The IGN, DGFiP and INSEE databases also differed in terms of their geographic positioning: tax sources allow for geocoding in the centre of (cadastral) parcels, whereas the IGN database overwhelmingly performed its geocoding along roads, in line with its road map layer, while the INSEE database recorded the position of addresses within its own application (called CICN2); its geometry was not automatically consistent with that used by the IGN. One particular issue concerned addresses with multiple locations, particularly in rural municipalities, in which there were localities corresponding to what may be very large areas in which roads were not numbered or even named. Equally, a single geographic coordinate might correspond to several separate addresses, for example, because of a lack of accuracy.

The prototype database that was developed started by trying to match the information in INSEE's own (ACL) database and the IGN database in terms of the address and/or the geographic positioning and then the location from the IGN's database was adopted. Where this could not be done precisely enough then a match was sought between ACL and DGFiP addresses and in these cases the location of the DGFiP's database was adopted. When neither of the two databases (IGN and DGFiP) provided a sufficient match, then ACL addresses were located manually by INSEE's regional offices.

The second step for producing the prototype database was to extend the coverage beyond that of INSEE's ACL database. Addresses and geocoding for municipalities with fewer than 10 000 inhabitants were taken over from the DGFiP database. For other municipalities, roads in the DGFiP that were not in the ACL database were analysed, checking among other things for possible mismatches (to avoid adding a duplicate).

A final step in this part of the project was to test the information in the prototype register, by using it to geocode the addresses of businesses held in INSEE's statistical business register (SIRENE). In 62.4 % of cases a business address could be geocoded directly, in 7.7 % of cases its position on a (known) road was estimated based on the position of the closest known house number, in 20.6 % of cases it was located in the middle of a (known) road and in 9.3 % of cases the address was located in the centre of the municipality (if the road was unknown). Approximately 100 businesses in three very small municipalities historically not on cadastral plans could not be geocoded. For a sample of businesses, a comparison of INSEE's statistical business register was made with the IGN database, which resulted in a slightly higher frequency of geocoding based on the address, possibly reflecting more information on additional roads or a better road recognition algorithm; this higher frequency may also (potentially) be related to a higher incorrect geocoding rate but this would require more analysis to be verified. This comparison also showed that geocoding was particularly difficult for agricultural holdings, mining and quarrying units (all of which tend to be concentrated in rural environments, where roads are more frequently unnumbered or even unnamed) and for public administration entities (as these are less likely to have been geocoded coherently as there is little interest in such entities for tax purposes).

Figure 2: Proposed organisation for INSEE's common address register



Action 3: as part of a longer term and sustainable strategy INSEE also worked on the development of an in-house common address register (RCA), with a common infrastructure. Initially this was designed to include residential addresses in large municipalities for the purpose of collecting census information. Geocoding for 85 % of these addresses come from the IGN with the remainder based on tax sources or manual location by INSEE regional offices. It was proposed to extend this strategy to all addresses, not just those used for the population and housing census. Furthermore, a suggestion was made whereby the information used for topographic base maps should also be combined within this action.

INSEE initiated a questionnaire for statistical authorities within various ministries in order to identify their needs and expectations with respect to managing addresses and also to identify their involvement with the national address database (BAN) project.

RESULTS

Action 1: INSEE's ACL is a geocoded address database updated on an annual basis in cooperation with municipalities; however, it is limited to residential addresses in large municipalities with over 10 000 inhabitants.

The French tax authorities possess a dataset describing premises which, combined with the cadastral map, can be used to compile a geocoded address register. These addresses are positioned in the centre of parcels, which does not always allow for them to be identified by a precise location.

The national mapping agency provides free access, for administrations and institutions with a public service mission (including INSEE), to a geocoded address database whose data is mainly sourced from tax files enhanced with data provided by various other partners. Although there are regional variations in the quality of address positioning, which is especially poor in rural areas, the IGN carries out geometric processing which is capable of refining address positions while ensuring overall consistency with its other map resources.

OpenStreetMap (OSM) France launched a database (called BANO) in 2014. It includes addresses from contributors (based on a crowdsourcing principle) and from the tax authority's cadastral vector map, with a list of roads listed by tax authority and local authority in open data format. The association seeks help from its contributors to harmonise addresses and roads from these different sources. However, BANO remains an incomplete database and national coverage is uneven.

Many administrations and public services use address data on a daily basis to manage registers for their specific professional activities. These addresses are rarely geocoded at source and are more generally entered freely without any normative framework. As a consequence, substantial efforts would be required for geocoding this information.

The first comprehensive, nationwide address database in France (base adresse nationale; BAN) was officially launched in April 2015 and aims to ensure wide-ranging public involvement (of administrations, public corporations and citizens) in the creation and maintenance of a high-quality national address database. Discussions have taken place in workshops involving numerous users and potential contributors, with a particular focus on developing a data model to improve the database structure in response to different (user) needs.

Action 2: the creation of an address register by merging various information sources requires investment in an efficient matching engine for the wording of addresses, which incorporates the common variations in address formulations.

There are discrepancies in the geographical positioning of addresses in the IGN, DGFiP and INSEE databases, which relate to the type of location method used. A non-standardised address may be difficult to geocode precisely on this basis alone if addressing efforts are not undertaken locally.

The development of an algorithm for the creation of an address register by merging different sources requires a form of calibration with a view to reducing the risks of omitting and duplicating addresses. In certain cases, monitoring the plausibility of matches among the addresses in several databases may require the inclusion of a manual validation stage. Depending on the desired level of precision, when none of the sources are able to provide a precise geographical location, positioning carried out on the ground may prove to be necessary. The initial ACL contained 5.7 million addresses, to which were added 687 thousand addresses for roads already existing in the ACL, 179 thousand addresses for new roads created in the ACL (in municipalities that already existed in the ACL) and 15.6 million addresses for small municipalities.

The initial results of a geocoding test for the statistical business register indicated that the processing carried out by IGN was more efficient, which is partly explained by the fact that its address register is more exhaustive in business sectors that are not so thoroughly referenced by the tax sources. However, even with the IGN system there remain a significant number of establishments for which precise geocoding does not exist.

Action 3: almost all of the ministerial statistical authorities in France are interested in geolocation, mainly for the production of statistics or studies/analyses for local territories. Several authorities also expressed interest in indicators of proximity and access times (for questions of access to cultural establishments for example). In the short term, the BAN is not able to meet the main expectations of official statistics, but official statistics could consider feeding into the BAN and thus enable a greater degree of data sharing and better convergence of address systems.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!QP34pq
INSEE: https://www.insee.fr/en/accueil
Statistical office's maps: https://www.insee.fr/fr/information/3287519
Mapping portal: https://statistiques-locales.insee.fr/#c=home

Croatia

Croatian Bureau of Statistics

CROATIAN BUREAU OF STATISTICS

Merging statistics and geospatial information in Member States — integration of spatial information into the statistical business register; 2014 project; final report December 2016

KEYWORDS: grid, business register

PROBLEM

Compliance with Regulation (EC) No 177/2008 requires each unit in the statistical business register to be georeferenced to a precise point on a map. However, the register in Croatia only included basic parts of addresses for local units and enterprises and was missing information on geographical location.

OBJECTIVES

The aim of this project was to add information on geographical locations (geo-referencing) to the statistical business register (SBR) held by the Croatian statistical office. In particular the objectives were to:

- upgrade the SBR with geocodes for legal and local units, and indirectly for enterprises;
- connect the SBR with the spatial statistical register (SSR) through a geo-reference code and the address for each legal and local unit;
- establish internal processes required for a continuous automated update (at least once a year) of geo-referenced data;
- publish a geographical presentation of selected SBR data on the statistical office's website.

METHOD

Action 1: development of the application for coding streets, upgrading the SBR with new attributes and creating links with the SSR application, retrieving geocodes from the SSR database and assigning them to addresses.

The key part of this project was related to information technology. As the resources were not available in-house for the necessary developments this was outsourced. The main elements included the following.

- The development of an application for coding streets based on a thesaurus of street names as well as maintaining this thesaurus to update it with new variations of street names. An initial analysis of the use of non-standardised street names was made and then a set of standardised street names was developed as well as street codes. Through automatic or manual (see below) procedures, these standardised street names and street codes were implemented in the register.
- The addition of new attributes within the address data in the SBR database and within the SBR user interface. An extension of the existing interface and underlying database made it possible to include more information (notably for the street code, NUTS region and geographical location, as well as more address information) from the SSR in the SBR. These structural changes to the database and interface were implemented not only for the live version of the SBR but also for tables containing historical information on addresses.
- Linking the new application with a live version of the SBR database and the new SSR database and the creation of procedures for automatic coding of streets and the assignment of geocodes to addresses. As such, the linking of information can be done in real time, with matched data taken over from the SSR to the SBR.

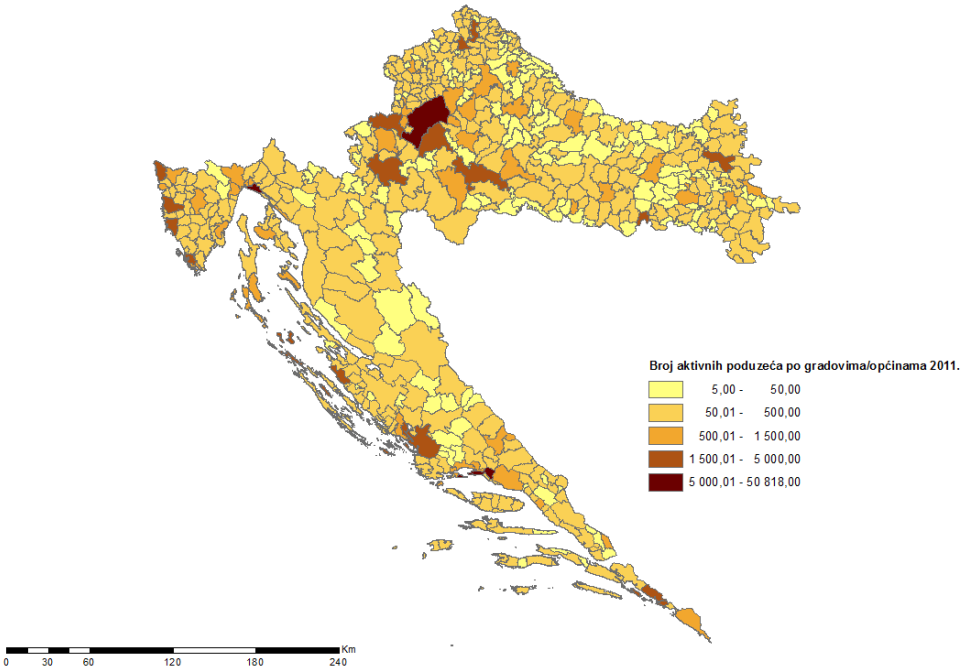
Action 2: manual matching of street names from the SBR with official street names in the SSR and populating the thesaurus.

The automatic coding of streets resulted in very low percentage of coded streets (only 40 %) and a lot of manual work was needed to match street names. The problems encountered were:

- a street name did not belong to the settlement where the unit was registered, but was instead located in a neighbouring (usually larger) town or village;

- a street name was 'old' — after the registration of the unit in the SBR, local authorities changed the name of the street and this was not reflected in the SBR;
- the administrative source had misclassified a settlement due to the fact that there are places with the same name but in different municipalities or counties;
- the street did not exist in the SSR due to the fact that the register of territorial units had not been updated or was not consistent with the real situation, for example when local authorities made changes but they were not reported (in time) to the state geodetic administration (SGA).

Figure 1: Number of active enterprises in cities / municipalities in 2011



Due to the fact that an exhaustive historical list of changes to street names is not available from the SGA, several cadastral offices in larger cities were contacted in order to obtain lists of street names that had changed, in order to be able to assign codes to streets that were not found in registers. Due to the difficulties encountered, manual matching was not employed for dead units in the SBR.

Action 3: selection and preparation of SBR data for publishing on the Croatian Bureau of Statistics (CBS) geoportal.

The only statistical output from the SBR was business demography data and these were used for geographical presentation. In 2015, a GeoSTAT portal was developed and this provides the possibility to publish data down to the level of 1 km^2 grids. GeoSTAT was also used for the presentation of SBR business demography data at NUTS level 3.

RESULTS

Prior to the work on this project, the SBR contained address data (postal information) and codes for three administrative divisions (at the level of counties, municipalities and settlements). As a result of the work done this was extended to include information on geographic location, a street code and codes for statistical divisions (NUTS levels 2 and 3). Furthermore, the quality of address data in the SBR was improved greatly as many mistakes were corrected. A system was established that should help maintain the improved quality of data on addresses by identifying problematic data at the moment that data entry takes place.

The published business demography data concerned the number of active enterprises, enterprise births and enterprise deaths, for two years, for each section of the Croatian activity classification (NKD2007); these data are available through the GeoSTAT portal.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!dQ64kD
Croatian Bureau of Statistics: https://www.dzs.hr/default_e.htm
Mapping portal: https://geostat.dzs.hr/?lang=en
State Geodetic administration: https://dgu.gov.hr/

Hungary

Hungarian Central Statistical Office

Merging statistics and geospatial information — merging address registers and distributing geocoded statistics; 2014 project; final report 21 December 2017

KEYWORDS: accommodation registers, visualisation

PROBLEM

After successfully identifying and joining the business register to an address register/directory, thereby integrating registers and geodata systems, two further economic registers were chosen for the same treatment — the register of public accommodation establishments and that for non-profit collective accommodation establishments. The Hungarian Central Statistical Office also identified a need to develop an interactive mapping application that could be used to disseminate its most important statistical indicators and their spatial distribution in a harmonised manner.

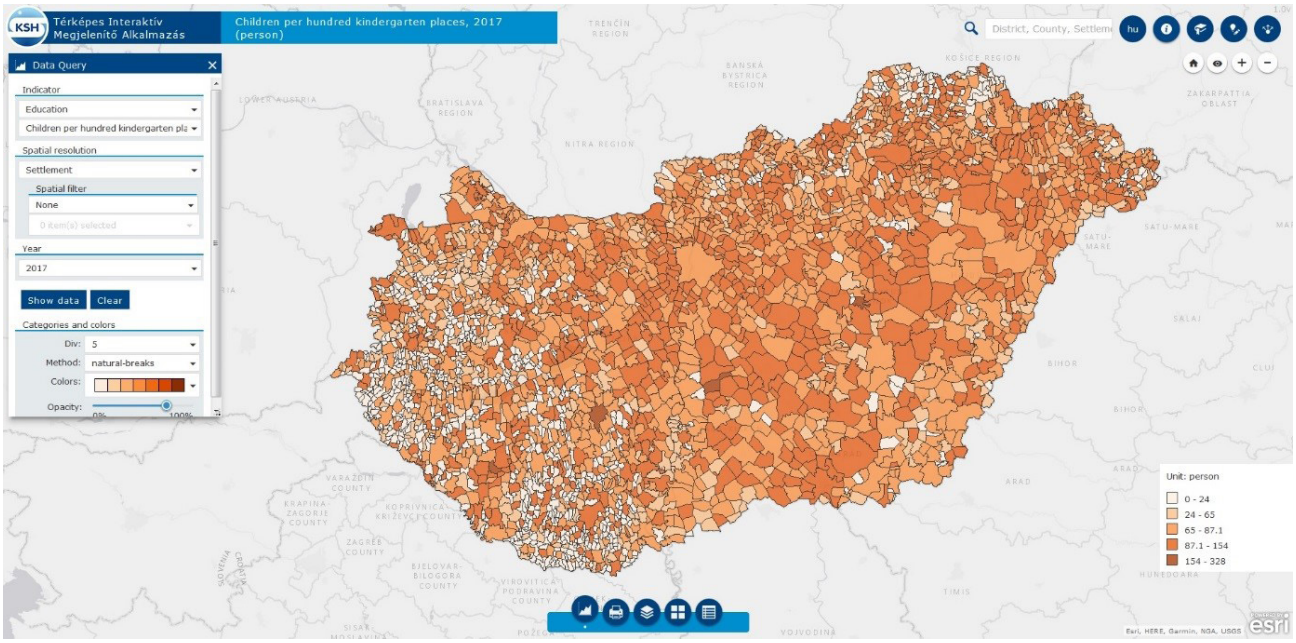
OBJECTIVES

Action 1: continue work on the preparation of the integration of various registers and a geodata-based system through the address register and address directory ([1]), thereby widening the scope of statistical data that can be merged with geospatial information. The two registers that were the subject of the work were the register of public accommodation establishments and the register of non-profit collective accommodation establishments.

Action 2: ensure public access to geocoded statistical data. The existing interactive mapping application on the statistical office's website had a number of disadvantages. In order to give more attention to geocoded data (grid-based information), the statistical office committed to develop a new interactive application to give supplementary information to users and offer solutions for user complaints that were identified.

([1]) For more information about the address register and address directory see the 2012 project presented on page 14.

Figure 1: Interactive mapping application



An interactive mapping application was developed whereby users may choose from a number of broad statistical themes (such as population, social or economic statistics), then select one of the available indicators for that theme and the desired territorial typology (regions, counties, districts or settlements, or grid cells for some indicators). The resulting map can then be personalised by determining the class boundaries and colour scheme. Users have standard navigation tools (zoom in/out and pan) as well as being able to mouse over any polygon of a statistical division in order to view the code and name of the area and its value for the selected indicator.

METHOD

Action 1: addresses for a registry of public and non-profit accommodation establishments (as well as those from some other registers) were stored and managed in a registry of public areas (known as the F054_KÖZTER table).

The first step of the work was to establish a connection between this registry of public areas and an address register and address directory.

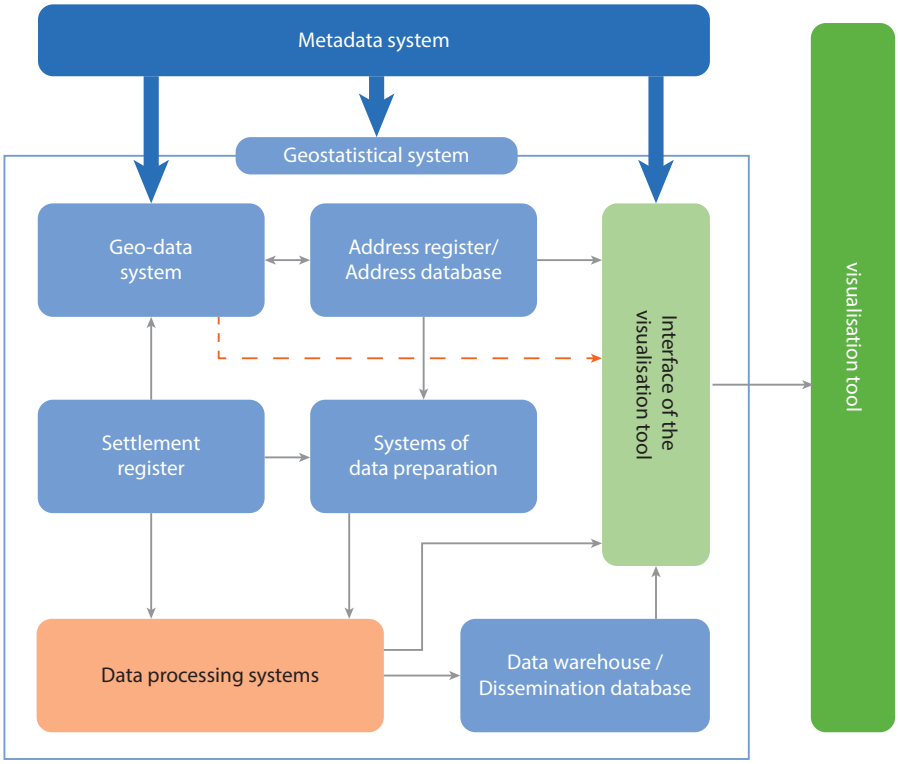
An existing address maintenance application was updated. The addresses in the register of public places were checked and corrected automatically by making use of this application. To help the manual checking of questionable addresses a detailed guide was compiled. Possibly erroneous addresses were identified and checked using various alternate address databases, as well as Google maps and other sources of information. Remaining non-identifiable addresses were verified by more than 1 000 local government authorities.

The statistical registers to be linked to geocoded data were prepared for linking by adding new fields and modifying maintenance procedures.

Once the checking of addresses had been completed the registers were connected to the address register and address directory.

Action 2: best practices were identified and studied and a report was compiled for the required functions and services for any new geocoded system. Equally, possible platforms and software were studied and in some cases tested and then a solution was selected: hardware and software were acquired and a new geographic information system (GIS) webserver and operating system were installed. Training for these new information technology (IT) tools was undertaken. The statistical topics and their spatial distribution were defined, alongside other functionalities.

Figure 2: Logical data model of the planned geostatistical system



A geostatistical system was planned to set up the conceptual connection between the existing systems and the new interactive application, with the aim of producing detailed spatial data automatically and transmitting these data to different interactive tools. Until such a system is implemented, basic data were compiled manually and data protection (confidentiality treatment) performed.

The functions and performance of an interactive mapping application were tested. A technical description, maintenance documentation and a user manual were compiled.

RESULTS

A register of public accommodation establishments and a register of non-profit collective accommodation establishments were linked to the address register and address directory.

A new interactive mapping application was developed and a geostatistical system was planned.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!GW33Nx
User manual: https://circabc.europa.eu/sd/a/6b28b02b-d554-451b-ae0c-ec75d4b0a21f/08143.2014.002-2014.394_UserManual.pdf
Hungarian Central Statistical Office: http://www.ksh.hu/?lang=en
Mapping portal: http://www.ksh.hu/interactive_humaps?lang=en
Metadata for registers: http://www.ksh.hu/apps/meta.menu?p_lang=EN&p_menu_id=1410

Poland

Statistics Poland

Statistics Poland

Merging statistics and geospatial information in Member States — support decision-making processes by combining statistical data with spatial data; 2014 project; final report 30 December 2016

KEYWORDS: analytics, transport, education, decision making

PROBLEM

Three specific problem areas were identified, each of which was linked to combining statistical data and spatial data within the context of real world examples, highlighting areas where policymakers and decision makers could benefit from exploiting new forms of data. The first concerned an evaluation of changes in the number of school children and the demand for school places to assist local government planning. The second was a study analysing the distribution of economic activity in urban centres in relation to infrastructure developments such as multi-modal transport accessibility. The final study was designed to fill a gap by providing a tool to visualise statistical data using topographic data.

OBJECTIVES

Action 1: this part of the project was designed to develop a methodology for studying and visualising the extent to which places in primary schools in 2014 reflect the need for such places (whether there is an excess of available school places or an insufficiency) and to project the situation for 2020 taking into account the demographic situation. In addition, teaching conditions were examined in terms of differences in the average number of pupils per teacher. Local governments face the challenge of adapting their network of schools and local infrastructures to changes in numbers of children while respecting binding standards.

Action 2: aimed to develop a methodology for analysing the accessibility of selected economic activity centres to multi-modal transport and the distribution of economic entities (hereafter called businesses) in relation to certain infrastructure elements.

Action 3: aimed to develop a methodology for assessing the suitability of a database for topographic objects for visualising statistical data and the production of spatial statistics.

METHOD

Action 1: this action focused on data for the voivodship of Mazowieckie (the voivodship containing Warsaw). Poland has in recent years seen a reduction in its total number of births due to lower fertility and postponement of motherhood, directly resulting in a progressively smaller number of pupils in schools; in Mazowieckie a low point in the number of pupils was reached in 2010, after which the number started to increase again, at least in part influenced by the lowering of the age for the start of compulsory education. Demographic projections indicate that the number of pupils will continue to rise until a peak is reached around 2020 after which it will decline again.

The work was done in two stages, a pilot stage for a narrower geographical area during which the methodology was tested, followed by the implementation of the methods for the whole voivodship. The pilot study involved defining assumptions and the data required, obtaining, correcting and merging the data, calculating indicators and analysing the results. The following stage followed a similar path to that for pilot study, with updated data and verification rather than development of the assumptions and methods.

The following data sources were used: data from the Ministry of National Education for pupils/students, teachers, schools, school rooms and primary-school identification data; town hall and municipality (gmina) data for district schools; social study frame data for population data. Whereas data for Warsaw was available electronically in a uniform structure, the data for rural areas was more varied in structure and format and required standardisation; data from various sources were merged and gaps filled. The information on the boundaries of school districts (school catchment areas) were sometimes out of date and so not all school districts could be linked to a complete set of addresses. A range of other issues were identified where information in various data sets did not match, such

as outdated information where more than one municipality had a school district covering a particular locality, or outdated information concerning changes in the status of schools (moving from public to private management or simply closing).

The calculation of data for 2014 and the estimates for 2020 took account of a number of factors. To estimate the requirements in 2020 it was necessary to take into account changes planned for the starting age of education in the school year 2016/2017. Furthermore, based on 2014 data, a relationship was calculated between the number of children attending a given district school and the number of school-age children actually residing within this school's district. Another adjustment needed was for schools that had opened in 2014 or recent years: these schools had an atypical age structure of pupils as they mainly had children in the first school year(s), whereas by 2020 they could be expected to have children across all school years as each cohort moved on annually. In a similar way, schools whose district (catchment area) had changed were also treated separately.

As well as calculating the number of pupils and the availability of places, several indicators were compiled:

- number of classes (called branches in Polish) per room used for conducting lessons;
- number of pupils per teacher;
- percentage of children transported to school because their homes were too far from district schools.

Where these indicators showed atypical results, the underlying data were examined in detail to see if there were reasons for the results or possible errors.

The final step for this action was to produce a visualisation application for the indicators. The outlines of school districts were established for 2014 and 2020 by aggregating census districts to school districts; adjustments were made for known changes to school districts planned to have been implemented by 2020.

Action 2: the focus of this activity was on the cities of Białystok and Toruń which were target areas for a study of transportation. An analysis was conducted into the distribution of economic activity with reference to infrastructural and the multi-modal accessibility of transport (road, rail and tram-based). For analytical purposes, grid systems were used: assigning businesses to grid cells enabled information on these businesses and their characteristics (legal form, form of ownership, main type of activity) to be aggregated by location and also to relate them to the availability of infrastructure such as networks for transport, sewerage, water supply, district heating and gas distribution.

A preliminary step was to identify the functional areas for the two cities. The following indicators were developed:

- the share of hired labour commuting into Białystok and Toruń within the overall number of persons commuting to these cities from each gmina (municipality);
- the share of persons leaving particular gminas to work in Białystok and Toruń within the total number of people leaving those gminas for work;
- the share of de-registrations from Białystok and Toruń in the overall number of de-registrations from these cities by gmina of the present registration.

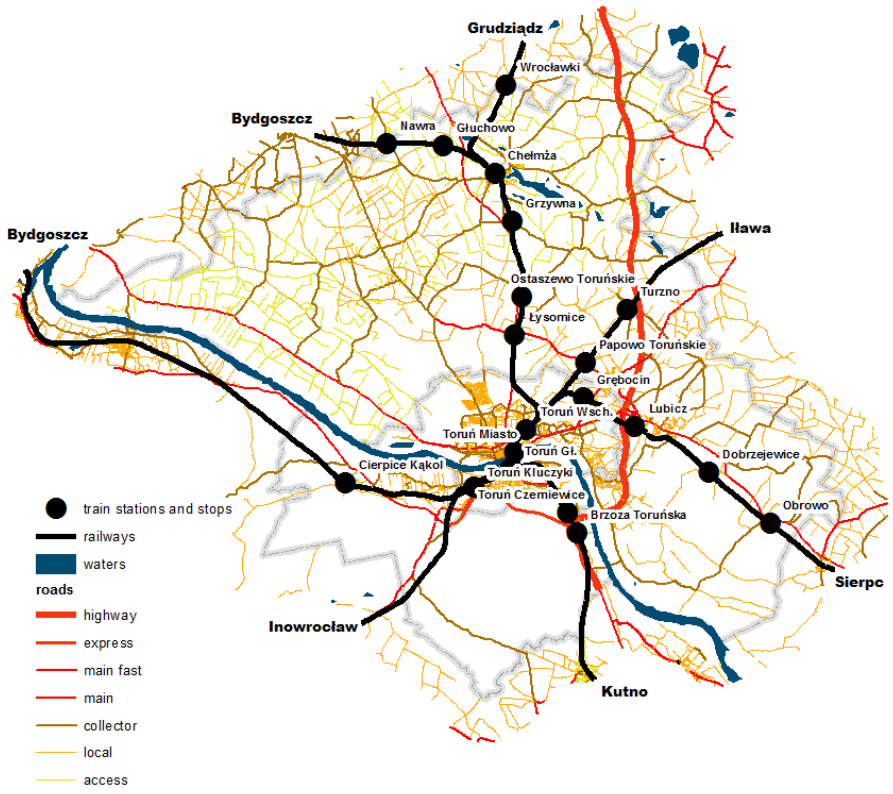
Based on this, two functional areas were identified, being composed of territorial units that had minimum values for all three indicators. This was done in such a way that the functional areas were spatially continuous, only containing territorial units that bordered each other.

The source of vector data on technical infrastructure was the database of topographic objects (BDOT) from the mapping agency (the head office of geodesy and cartography). An orthophotomap and a vector layer of roads and streets (DRUL) used in work related to the maintenance of spatial address databases (as part of the TERYT register), were applied to evaluate the degree of land cover with different objects. Spatial data about businesses were sourced from the statistical unit database (BJS) for the two cities concerned. Public transport connections (buses and trains) to the cities of Toruń and Białystok from their respective functional areas were acquired from Blue Ocean Business Consulting.

Three different analyses were then performed.

The data from BDOT was converted for use in the software adopted for this project and evaluated. For several of the networks (such as district heating, water supply or sewerage) the networks shown in the data were found to be largely incomplete and further investigation showed that these only concerned over-ground sections of the networks. Other sources were investigated but all required payment for establishing datasets. From 2018, information on complete networks are expected to be available in a national database and will be available free

Figure 1: Visualisation of the transport network



of charge to the statistical office. The completeness of data concerning rail and tram networks was evaluated and it was established that the data were complete; in practice, the data for the tram network were not used. Road data from BDOT were also evaluated and used. The data from the statistical unit database were analysed and it was decided to exclude a small number of businesses, such as those under bankruptcy or liquidation proceedings, as well as those related to private partnerships. The distribution of businesses according to legal form, ownership form, size and economic activity was analysed. For the presentation of data five economic activities were selected: manufacturing; construction; distributive trades; transportation and storage; and professional, scientific and technical activities. To compare the information from different sources for roads and businesses, the linear data for roads were converted into density data (kilometres per km^2) and the point data for businesses (and their characteristics such as levels of employment) were also converted into density data (ratios per km^2).

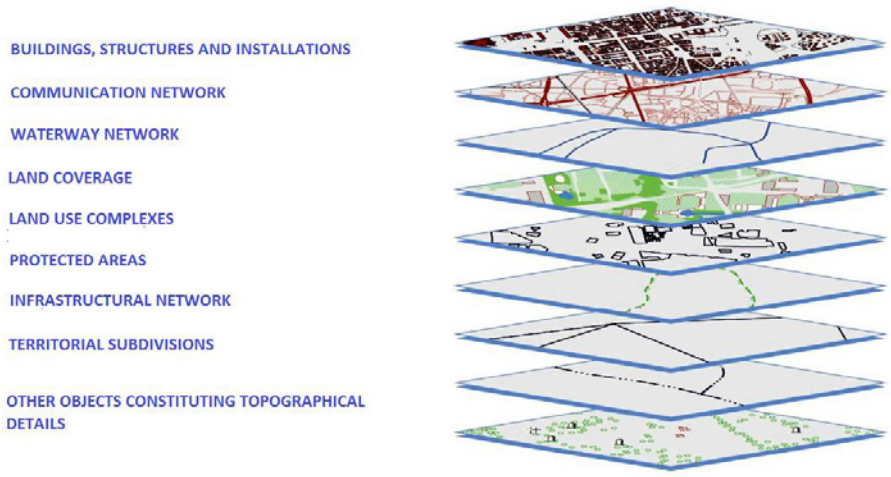
For the second analysis, which was based on transportation accessibility for bus and train connections, the focus was on access time, looking at time to and from the two cities as well as the number of connections between them. An analysis was made of bus and train timetables between (to and from) the two cities and the municipalities within their functional areas, looking at three two-hour intervals in the morning (between 4 and 10 a.m.) and in the afternoon/evening (between 2 and 8 p.m.). This analysis resulted in approximately 5 thousand connections (for the two cities combined), of which the vast majority were bus connections. For each locality within the functional areas, an index was then compiled of the average commuting time for each two-hour time interval, based on the total time for all connections and the total number of connections. These averages were then adjusted to allow for the closeness of the time intervals to the peak commuting time and also to allow for time intervals in which there were no connections. Finally, the index values were summed across all time intervals to produce an overall index for a locality. This index was compiled separately for buses and trains and as a joint index.

The third analysis focused just on the city of Toruń and analysed accessibility by road and by rail. The information from BDOT on rail tracks was filtered to exclude tram rails as well as rail tracks without passenger traffic, for example ones that were no longer used or were only used for freight. The remaining track information was supplemented by manually adding information on stations and stops. The railway tracks were then split up into segments between stations/stops and the travel time for each segment was added. The information from BDOT on roads was refined to include only usable road sections and again travel times were added to road segments based on the maximum speed limits and a weight for encountering traffic difficulties. The road and rail networks were then visualised for various road and rail classes.

Action 3: this final action focused on data from the Database of Topographic Objects (BDOT), initially based on 2012 data and then updated with 2015 data. BDOT data covered nine categories of object classes: water networks (natural and manmade); communication (transport) networks; energy infrastructure networks; land cover; buildings, structures and devices; land use complexes; protected areas; territorial divisions; other objects.

The second source studied in this final action was the land and building register (EGiB), an official register containing information on real estate. It contained information on land, buildings and premises and indicated (among other characteristics), the names and address of owners of real estate, the cadastral value of real estate, and information on lease arrangements. In vector form, the register did not completely cover all urban areas in the three voivodships or all of the rural areas in the 10 voivodships, while the descriptive part of the register was also incomplete.

Figure 2: Visualisation of feature class categories in BDOT10k



The final source was orthophotomaps.

The first step was to perform an analysis of data quality in the BDOT, looking at data for single-family houses which formed part of the buildings, structures and devices class in the dataset. The focus was on the Mazowieckie Voivodship, in other words the Polish capital city region, the largest NUTS level 2 region in Poland in terms of area and population.

The vector map objects from the BDOT could only be compared on the basis of an othophotomap, given the incomplete EGiB coverage. Evaluation was performed by way of selecting a sample comprising 10 % of grid squares. The sample selection process involved an earlier identification of the spatial diversification of residential buildings in the voivodship, followed by drawing fields divided into a few types of areas: city centres (6 % of the sample), suburban areas (47 % of the sample) and rural areas (47 % of the sample). Data were verified in terms of:

- completeness of objects entry;
- compliance of object boundaries with the orthophotomap;
- correctness of entered attribute values;
- correctness of the identification of objects.

The evaluation resulted in a report containing test field ranges and a list of errors.

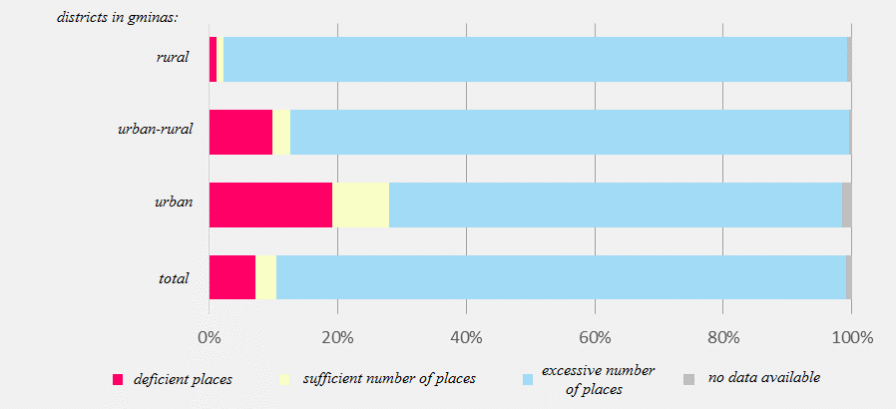
Based on the initial assessment of the quality of the BDOT data, the second step was to compile a spatial analysis of data selected from the source. Single family houses were selected: for each of these the location of the centroid was identified and the area added as an attribute.

RESULTS

Action 1: on average in 2014, one in eight schools had less school places than children (who should be in education) residing in a given district. The largest proportion of such schools was recorded around Warsaw. Conversely, an excess of school places (compared with the number of pupils) was recorded for most schools, with some having several times more places than school-age children: these were mostly in the north-eastern and southern part of the Voivodship.

An apparently insufficient number of places for children residing in a school district does not mean that these school places were not available: in fact, a real insufficiency was recorded for only half the schools where the number of children residing in a given district exceeded the number of places. A number of factors underlie this, one of which is the fact that parents are not obliged to send their children to schools in districts where they reside.

Figure 3: School districts by type of gmina and actual availability of places in 2014

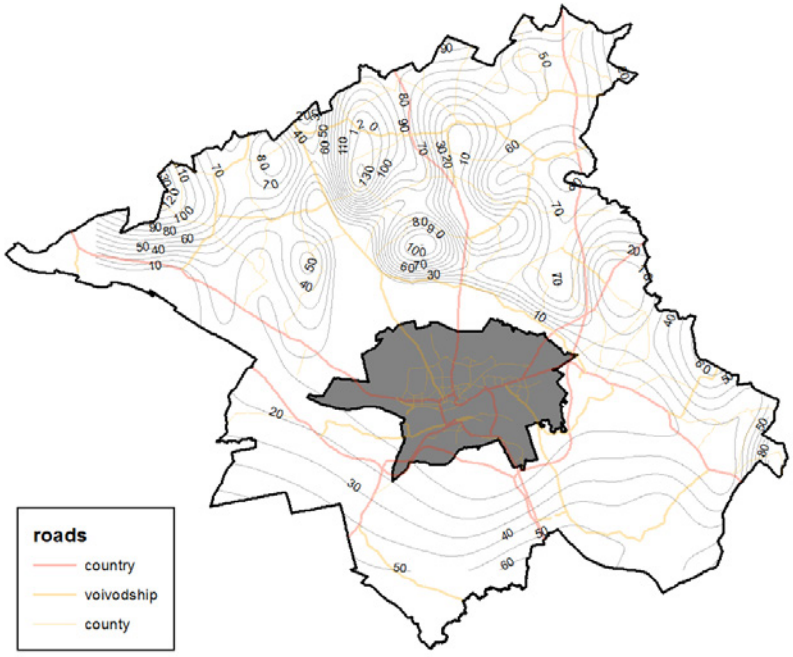


A large proportion of schools have an excess number of places. In many rural gminas (municipalities) the number of pupils per teacher was very low. In extreme cases there were an average of 2-3 pupils per teacher, with the average for all schools in the study being 11 pupils per teacher. Closing schools in rural communities often results in pupils having to travel further to school: in some gminas 90 % of children had to commute to schools and in one case the proportion was 100 %.

Concerning projections for 2020, the number of schools with insufficient places for children residing in school districts is expected to increase, with the situation likely to be most severe in the central part of the Voivodship, as well as in the north-eastern gmina of Łyse, in which the closure of four primary schools is expected to affect teaching conditions adversely. Conversely, the percentage of schools with an excess number of places is expected to increase mainly in southern parts of the Voivodship. Pupil transport issues are likely to increase as more and more schools face closure.

Action 2: the original intention when looking at business concentration was to compare this with the concentration of a range of networks, but in the end the only reliable and comprehensive comparison available was with the road network. Areas of the two cities were identified which had above average business density and areas with above average density for road networks and these were mapped, with various analyses for different types of businesses (by legal form, ownership form, size and economic activity). Overall there was little correlation between the location of businesses and the distribution of roads in either city although the mapping of the business data did show local variations depending on the different criteria used, particularly in Białystok. A further analysis was performed which identified where the higher density of businesses and the highest density of employment was in each of the two cities: while these concentrations (businesses and employment) were relatively close to each other in both cities — particularly in Toruń — they did not coincide exactly.

Figure 4: The accessibility from Toruń to the localities within the functional area between 2.00 p.m. and 8.00 p.m. according to the bus transport accessibility index (roads)



For accessibility by bus and train, maps were plotted showing either the road or rail network accompanied by isolines representing accessibility index values. These isolines ware not simply concentric circles around the city centres, but were curved to reflect on one hand the location of major roads and rail lines and on the other hand the frequency of rail services at different times of the day. Accessibility indices were plotted for the morning (going into cities) and for the afternoon/evening (leaving cities), which also made it possible to contrast these two situations and thereby identify locations with relatively balanced or unbalanced accessibility in these two directions.

For the final analysis, where road and rail networks were combined with information related to the travel time for each segment of these networks, it was possible to identify the most efficient (in terms of time) routes between locations,

for example between areas outside of the city into the city centre. When combining information for road and rail transport these identified the best of several nearby stations to drive to in order to take the train into the city in the shortest combined travelling time.

Action 3: no counterparts were found in the BDOT dataset for 68 objects (0.18 % of all identified objects) in city centres, 538 objects (0.45 %) in suburban areas and 202 objects (0.54%) in rural areas. Equally various attributes were examined, such as the number of storeys and the area and the proportion of errors was also regarded as low. It was concluded that the dataset fulfils the criterion of usefulness for statistical purposes.

Having prepared a dataset for single family houses the information on the number of houses and their area were summarised within grid cells and the results presented as density indicators within a choropleth map.

Figure 5: The most efficient travel method to Toruń from the north-western part of the functional area (based on the network model)

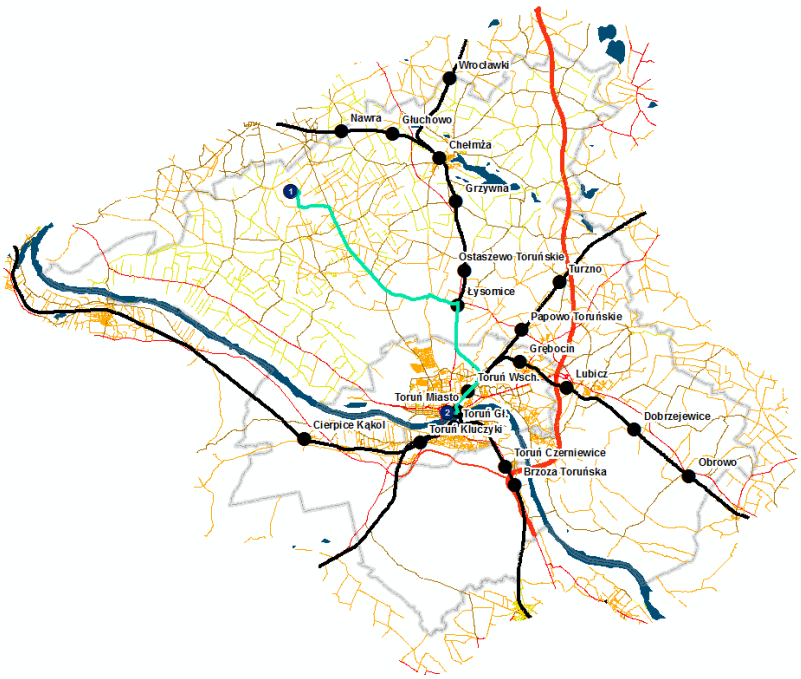
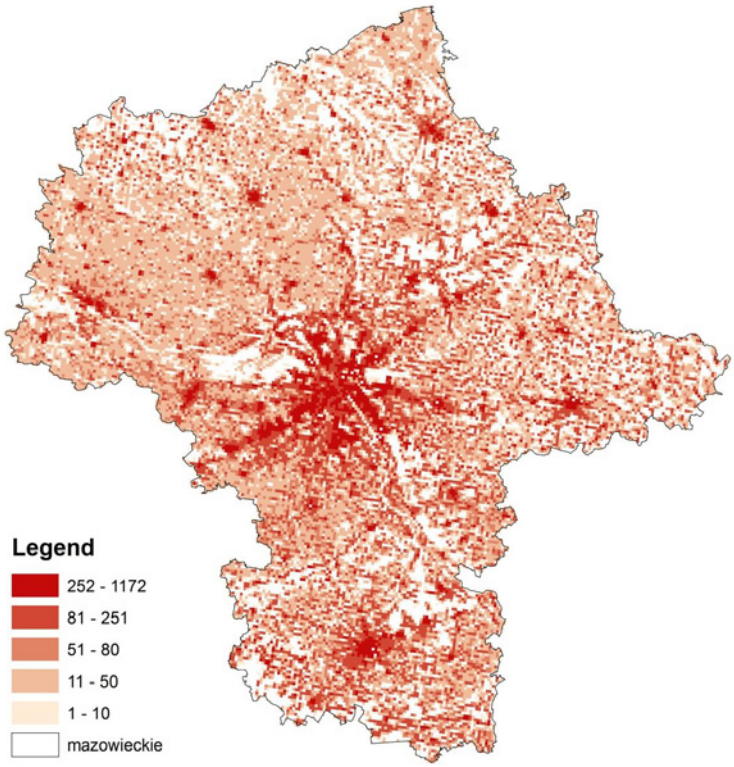


Figure 6: Number of single-family houses/ km²



FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!jM43fU
Statistics Poland: http://stat.gov.pl/en/
Mapping portal: http://geo.stat.gov.pl/en/

Portugal

Statistics Portugal

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Support policymaking by the use of spatial information combined with social, economic and environmental statistics; 2014 project; final reports July 2015, June 2016 and February 2017

KEYWORDS: business register, address

PROBLEM

Statistics Portugal (INE) identified a weakness in relation to their business register statistics insofar as these were not integrated into the point-based component of their in-house spatial data infrastructure, thereby impacting on the production of statistics for small areas and the availability of spatially-enabled datasets.

OBJECTIVES

The main aim of this project was to perform the harmonised geo-integration of statistics based on the statistical business register, by way of a spatially enabled and quality-controlled point-based infrastructure. The objective was to make use of: i) the existing spatial data infrastructure, available statistical resources, official registries and other administrative data sources; ii) the potential offered by geographical information technologies; and, iii) institutional relations with the National Mapping Agency (NMA) and municipalities.

The specific objectives were:

- to analyse the statistical business register concerning harmonisation of addresses and relevant attributes for geo-processing;
- to define methodologies and technical solutions for geo-processing and data integration;
- to implement, validate, promote and disseminate a business register geographic database.

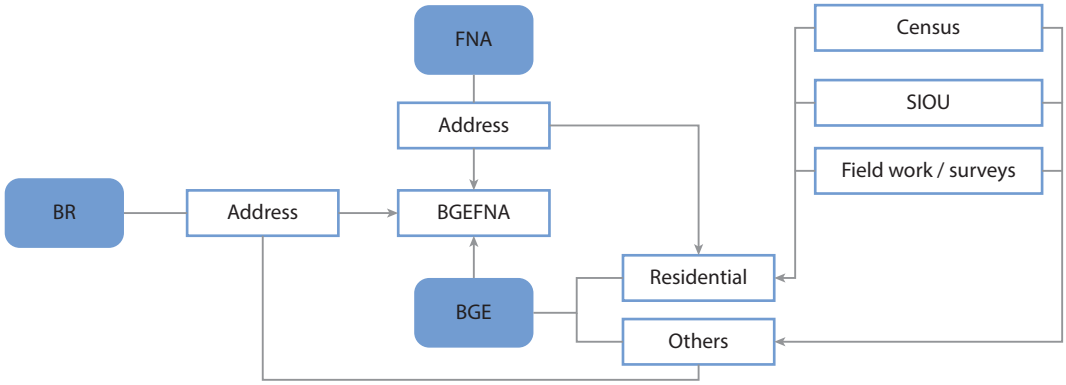
METHOD

Objectives 1 and 2: analysis and methodology.

The starting point of the work was to analyse:

- the statistical business register (FUE);
- the national dwellings register (FNA) — the master file to which any address based system should be matched for the purpose of geo-referencing existing databases within the statistical office;
- the buildings geographic database (BGE) — this has a point-based coverage of buildings that can be used for residence and has been enriched to record other types of buildings with no residential capabilities and is dynamically maintained by municipalities through the indicators system of urban operations (SIOU) with respect to building permits and completed construction work.

Figure 1: Inter-relationship of sources



Analyses of these elements was carried out with the perspective of preparing the design and production of a business register geographic database (BRGD).

The common attribute of the three main databases was their address component. The quality assessment of the address component of the statistical business register (FUE) revealed the need to further enhance and harmonise it.

The methodology for the BRGD was established, defining the processes for data integration, data flows for maintaining the BRGD, the data structure required, and the technical specifications for building the planned applications. Concerning data integration, a variety of issues were assessed, including, among others, a harmonisation procedure for addresses and other required attributes, and a process for dealing with multipurpose buildings. The technical specification included, among others, the system architecture, main structure and specific data model for the BRGD, as well as the validation platform for municipalities to verify the accuracy and the completeness of the initial BRGD and tools for internal visualisation and for the external geoportal, with spatial query capabilities. The proposed applications were designed in accordance with existing technological resources in order to benefit from a common infrastructure supporting the BRGD, the validation platform and the dissemination applications, as well as the entire in-house spatial data infrastructure.

Objective 3: implementation, validation, promotion and dissemination.

The steps in this objective were to:

- build a partial test version of the BRGD following the methodology developed under the first action;
- develop a validation platform and perform the validation;
- develop an internal application for dissemination;
- develop an external application for dissemination.

Three target subsets of units were selected to test filling the BRGD: the units needed for the consumer price index sample; all units (regardless of activity) in a particular region (Oeste); and all units (regardless of region) in a particular activity (Subclass 68 322). The main problems faced in matching units were related to the quality of the address, for example because it was wrong, had changed, had been misspelled, was in a business area, or was written differently — various approaches were implemented to improve the matching process.

A validation platform (GeoBGE) was developed as an extranet application with restricted access rights (permission). Various cartographic elements from multiple sources were integrated into the application as basemaps (for example Bing maps and Open Street map data) to facilitate the location and recognition of the geography within different perspectives for the territory.

The validation of the results obtained through the test implementation step relies on the long-term effective relationship between the statistical office and municipalities to make use of their extensive knowledge of local entities. The validation was done, using the GeoBGE platform, for one municipality (Torres Vedras) within the Oeste region. The municipality was asked to validate records where a match had been achieved, looking specifically at the location (point coordinates), address and economic activity of each record. Proposals by the municipality to make changes were then checked by the statistical office.

The development of the internal application (Geoplaneamento) was designed to assist the work of the methodology unit and the data collection and analysis unit within the statistical office. The internal application comprises not only administrative and statistical units (such as small statistical areas, NUTS, the location of the surveyors), but also includes the primary sampling units for samples referenced to households in a point-based approach supported by the BRGD. Some additional features, such as viewing Google Street View images or locators for household codes and addresses are included to help identify households.

The external applications aimed to give users direct access to statistical information. An example was the GeoEscolas application which used BRGD records for schools. The application provided information related to the surrounding area of a specific primary or secondary school. Spatial queries can be made of a 200, 500, 1 000 or 2 000 metre radius to obtain statistics related to buildings, dwellings, families, population (total and by age group).

RESULTS

Objectives 1 and 2:

The analyses and methodological developments related to data integration gathered all the relevant information to launch the activity to implement the development of the BRGD. Furthermore, the specifications for various applications were defined.

Objective 3:

It was concluded that the statistical business register (FUE) requires further harmonisation for address components; its accuracy might also be improved through the process of building the BRGD. The efficiency of the complete address as the locator to match records was enhanced by use of a common structure created to relate to the various datasets. However, the complete address is insufficient for geo-referencing the statistical business register (FUE) and a step-by-step approach based on elements of the address was valuable to assist and complete the process. Experience was gained assessing the usefulness of various locators to match records.

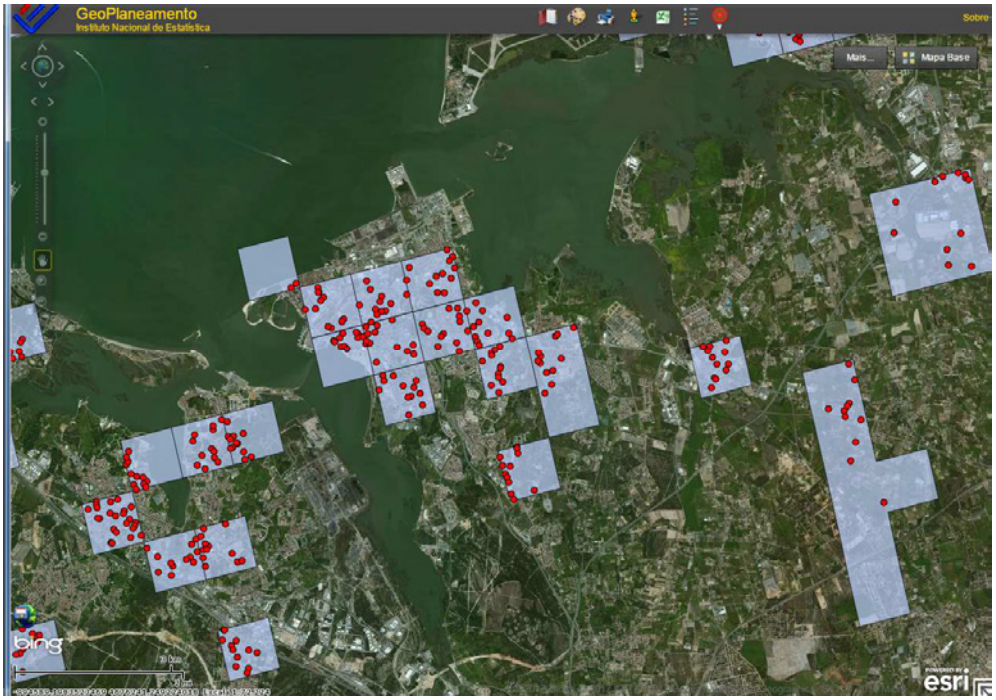
The complete implementation of BRGD requires prior investment to improve the quality of addresses in the statistical business register (FUE). The statistical office adopted an internal regulation specifying the data model to be used for addresses in all databases and systems, the purpose of which is to improve the quality of addresses in the business register (FUE) and the national dwellings register (FNA).

The validation platform (GeoBGE) created for municipalities to validate the BRGD is a mapcentric fully interactive web application. Users can navigate by zooming in and out or dragging. They can identify data by querying the attributes of specific geographic objects. GeoBGE is customised for data editing and validation.

For the municipality of Torres Vedras, 90.3 % of records were validated with no changes, while for 0.2 % of the records changes were proposed to all three attributes (location, address and activity); the remaining 9.5 % of records had changes proposed for one or two of the attributes, most commonly concerning a change to their address. Overall, 98.0 % of matched records were confirmed in terms of their location, 91.0 % in terms of their address and 98.6 % in terms of the economic activity.

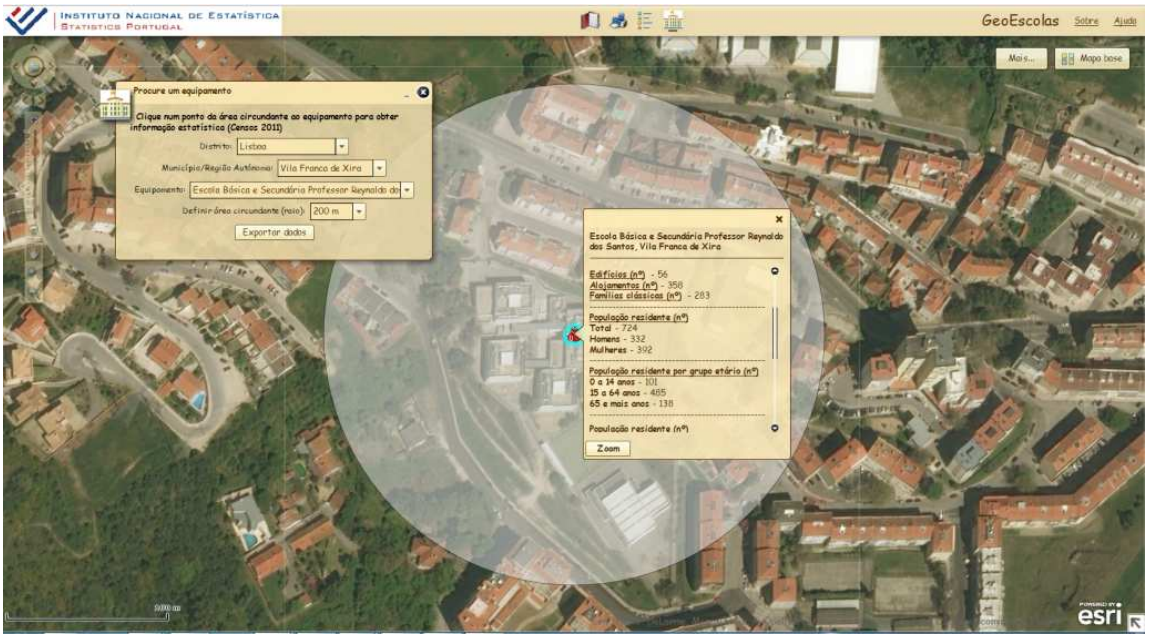
Geoplaneamento has been used by the data collection unit as it provides the exact location of those dwellings that were surveyed, supporting more efficient fieldwork planning and the identification of the closest surveyors available to collect data for specific dwellings. The methodology unit uses Geoplaneamento to analyse the dispersion and location of samples and the sampling frame for each survey.

Figure 2: GeoPlaneamento



The external web application was developed to be used by teachers and students and is freely available through the statistical office's website to promote statistical literacy.

Figure 3: GeoPlaneamento



FURTHER INFORMATION AND LINKS

Final report — analysis and methodology: https://europa.eu/!uj38fP
Final report — implementation: https://europa.eu/!xB33Pq
Final report — validation and dissemination: https://europa.eu/!cG66Dn
Statistical office: https://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE
2011 census (mapping) portal: http://mapas.ine.pt/map.phtml
GeoEscolas: http://geoescolas.ine.pt/

Norway

Statistics Norway

Statistisk sentralbyrå
Statistics Norway

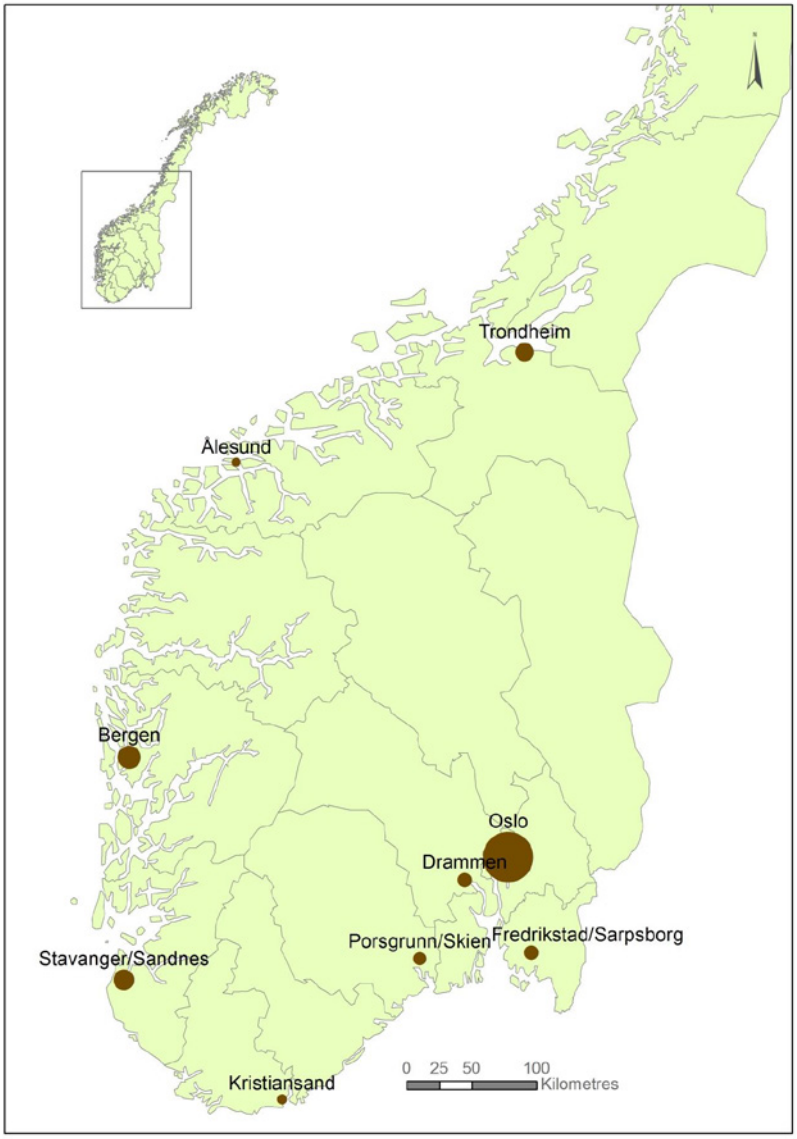
Mapping attractive urban areas — a way to geographically determine quality of life parameters of importance; 2014 project; final report 28 February 2017

KEYWORDS: analytics, quality of life

PROBLEM

Quality of life perception surveys can be used to determine attractive urban areas. However, they are a relatively expensive and time consuming task and generally result in aggregated data for the whole of a city, rather than more detailed information on specific neighbourhoods within a city. Statisticians have considered testing alternative data sources — such as statistical registers and georeferenced data — to see if these can be used to make the information gathering process more effective and efficient.

Figure 1: Location of the urban settlements in the project.



Note: the area of each circle is proportional to polulation size

OBJECTIVES

The general objective of this project was to combine relevant statistical registers and geo-referenced data in order to identify attractive urban areas. Its aim was to develop an innovative procedure for assessing how changes in population and land use in urban settlements relate to quality of life parameters.

METHOD

The first step was to describe the quality of data sources and the possibilities for combining these into a dataset for measuring the attractiveness of urban areas. From this a conceptual model of the data structure and data format for a dataset of urban area attractiveness was developed.

A methodology for producing a dataset on the attractiveness of urban settlements was developed. Housing prices (total price and price per m^2) of sales in 2014 were used as proxies for attractiveness, as they reflect supply and demand and are a numerical representation of some kind of attractiveness. Intrinsic characteristics of a dwelling such as its area, condition and so on also influence price but are less likely to determine a neighbourhood's attractiveness.

This project investigated whether there is a variation in house sales prices dependent on location within (not between) nine Norwegian cities, seeking to explain this by correlating price and place with factors related to the dwelling and its location.

In total, six categories of variables were considered: the dwelling itself; distance to geographic entities (such as the city centre, recreational areas, or a coastline); distance to buildings often providing services (such as schools, universities, or restaurants); intensity-environment (such as noise levels or hours of sun); population characteristics within 250 m radius (such as average income, education level, migration); and employment (number of employees within a 5 or 10 km radius). The indicators relating to a dwelling's sales price, floor space (in m^2) and age were available from real estate sales data, which were then combined with data from the geo-referenced property register in the Cadastre to get the XY coordinates of the centroid for each property. A large range of sources were used to provide data for the variables covering the five categories related to the location (rather than the dwelling itself).

In general, migration within a city was problematic as an indicator for measuring attractiveness, as the availability of housing in an area is quite often more related to urban planning issues than the area's attractiveness. New dwellings are not necessarily built in the city's most attractive areas, for example due to space issues and urban planners may wish to offer more affordable housing. Equally, new building permits are not necessarily located in the areas of a city that are perceived to be the most attractive. As such, migration and building permits were not selected as explanatory variables.

An ordinary least squares regression analysis was performed for the two dependent variables (total price and price per m^2) and the adjusted R-squared calculated for each of the explanatory variables that were to be tested.

RESULTS

The results of the analysis showed that the explanatory variables (other than those related to the dwelling itself) explained slightly more of the variation in total house prices than the price per m^2. Furthermore, they explained more of the variation in prices for cities of at least 150 000 inhabitants than for cities with fewer inhabitants.

Education levels and household income were found to be strong indicators of variations within cities for the total price of a dwelling. These were particularly important in Oslo indicating that it is more socioeconomically divided than other cities. Equally, floor space (in m^2) was also found to be a strong indicator of total price variation. Five other variables were found to be significant for determining total prices in Oslo, but to a lesser extent: the distance to restaurants, the city centre and to water, the mean age of the population, and the age of a building (in years or those dwelling simply built before the Second World War). No other variables (that were tested) were found to be significantly important in Oslo for the total price variation. Notably, the three perception survey variables (schools and higher education establishments, health services and public transport) were not significant and similar results were found for these variables in the other eight cities studied; it appears that the distance to these services is generally short enough throughout the cities that it does not significantly affect total sales prices. Equally, access to recreational areas, noise levels and employment opportunities were not significant within individual cities.

Table 1: Total sales sums — how much of price variation we are able to explain in Norways 9 largest cities. AdjR2 for each variable isolated, and total combined AdjR2.
(AdjR2 of 1 = 100 %)

	Urban settlement								
	Oslo	Bergen	Stavanger	Trondheim	Drammen	Fredrikstad	Skien	Kristiansand	Ålesund
Population	958 378	250 420	210 874	175 068	113 534	108 636	91 737	60 583	50 917
RESTAURANT- DISTANCE	0.00	0.00	-0.01	-	0.01	0.01	-	0.00	0.02
CITY CENTRE- DISTANCE	0.00	0.02	-	-	-	0.01	-	-	0.14
WATER-DISTANCE	0.01	0.01	-	0.00	0.00	-	0.01	0.03	0.02
FLOOR SPACE M^2	0.60	0.62	0.67	0.64	0.53	0.45	0.48	0.45	0.53
EDUCATION LEVELS - POPULATION	0.20	0.09	0.04	0.10	0.23	0.14	0.14	0.18	0.08
HOUSEHOLD INCOME - POPULATION	0.39	0.23	0.26	0.24	0.32	0.25	0.23	0.22	0.24
AGE – MEAN OF POPULATION	0.02	0.00	-	0.00	0.00	0.00	-	-	0.01
BUILDING AGE	0.02	0.04	0.00	0.00	0.05	0.04	0.06	0.08	0.01
COMBINED	0.82	0.77	0.75	0.79	0.74	0.58	0.57	0.70	0.64

Reading note: the table specifies how much each variable can explain of the variation in the total sales prices of dwellings; an AdjR2 (adjusted R-squared) of 1 is 100 %. The table also shows how much of the price can be explained by all variables combined, for example in Oslo all of the variables together explained 82 % of the variation in price. Among all of the variables that were tested the table shows the eight that were significant and consistently contributed to price variations in the expected direction in Oslo. The colour of the values in the table indicates whether the eight variables behave the same way in the other eight cities. Type 1 variables (values shown in black) are significant and contribute consistently in the same direction to price variations. Type 2 variables (values shown in blue) contribute in the same main direction, but not consistently or not in a significant manner. Type 3 variables (values shown in brown) are those which contribute in the opposite direction to that which was expected.

For the price per m^2, income level was not found to be a significant variable, whereas floor space and education level remained significant. There was a correlation between price per m^2 and variables related to distance, notably the distance to the city centre as well as the distance to restaurants (reflecting local centres) and higher education establishments (although many are centrally located). However, the distance to hospitals was not conclusive as a measure of attractiveness, suggesting that they are not necessarily perceived as an attractive neighbour and/or that they are not located in attractive areas.

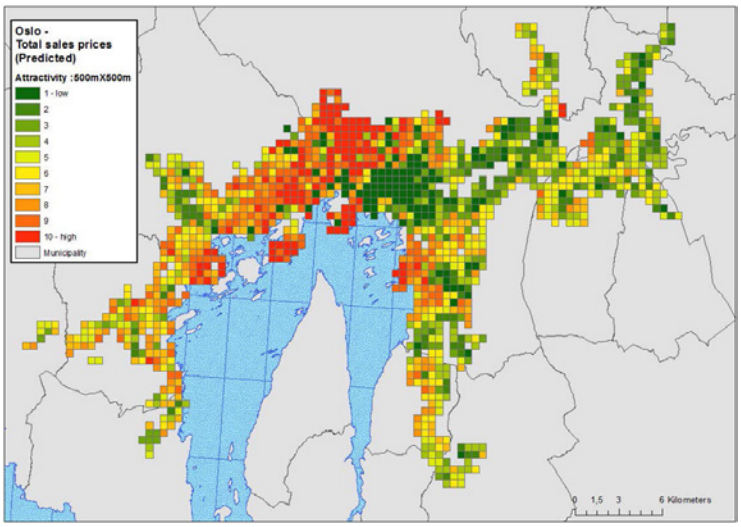
Table 2: Price per m^2 — how much of price variation we are able to explain in Norways 9 largest cities. AdjR2 for each variable isolated, and total combined AdjR2.
(AdjR2 of 1 = 100 %)

	Urban settlement								
	Oslo	Bergen	Stavanger	Trondheim	Drammen	Fredrikstad	Skien	Kristiansand	Ålesund
Population	958 378	250 420	210 874	175 068	113 534	108 636	91 737	60583.00	50917.00
HOSPITAL - DISTANCE	0.30	0.17	0.01	0.18	0.20	0.09	0.01	0.20	0.08
RESTAURANT- DISTANCE	0.24	0.21	0.01	0.17	0.17	0.05	0.08	0.31	0.13
EDUCATION LEVELS - POPULATION	0.35	0.30	0.06	0.10	0.03	0.01	0.00	0.05	-
CITY CENTRE- DISTANCE	0.43	0.30	0.04	0.35	0.14	0.03	-	0.30	0.02
WATER-DISTANCE	0.04	0.03	0.04	0.14	0.08	0.03	-	0.03	-
FLOOR SPACE M2	0.32	0.52	0.66	0.62	0.39	0.37	0.28	0.42	0.37
AGE – MEAN OF POPULATION	0.01	0.00	0.00	0.01	0.15	0.14	0.08	0.19	0.07
HIGHER EDUCATION - DISTANCE	0.19	0.27	0.02	0.12	0.13	0.02	0.03	0.23	0.02
BUILDING AGE	0.16	0.10	0.13	0.12	0.11	0.06	0.05	0.09	0.00
COMBINED	0.74	0.73	0.75	0.77	0.67	0.52	0.45	0.70	0.51

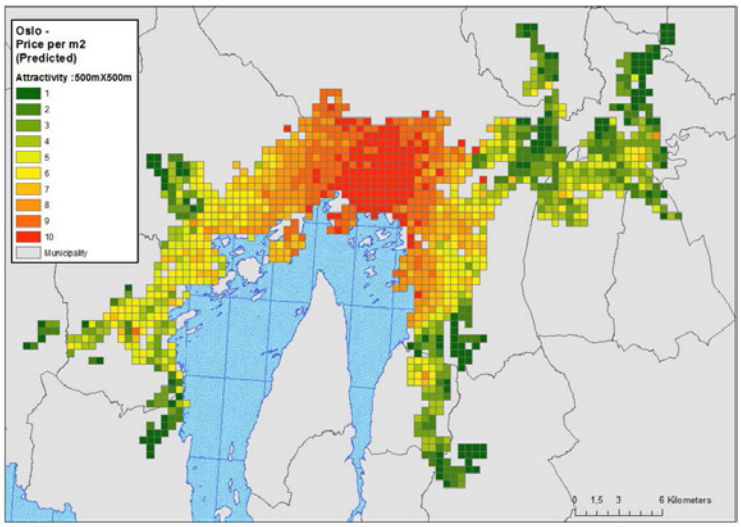
Reading note: see also the reading note for Table 1 above. Among all of the variables that were tested the table shows the nine that were significant and consistently contributed to price variations in the expected direction across Oslo.

Based on a regression analysis, two attractiveness indices were calculated for several cities, one based on the total sales price and the other based on the price per m^2. This was done by combining the coefficients calculated from data on house sales with data for the full housing stock in these cities using the geo-referenced building register in the Cadastre. The predicted data for the individual buildings were then averaged within 500 m x 500 m grid cells to classify each grid cell within a price decile: the top decile shows the 10 % of grid cells with the highest average predicted prices (total or per m^2) and the bottom decile the 10 % of grid cells with the lowest average predicted prices. The results for Oslo are shown in Figure 2.

Figure 2: Production of attractive urban areas



Predicted
Total sales price
Attractivity index 1-10



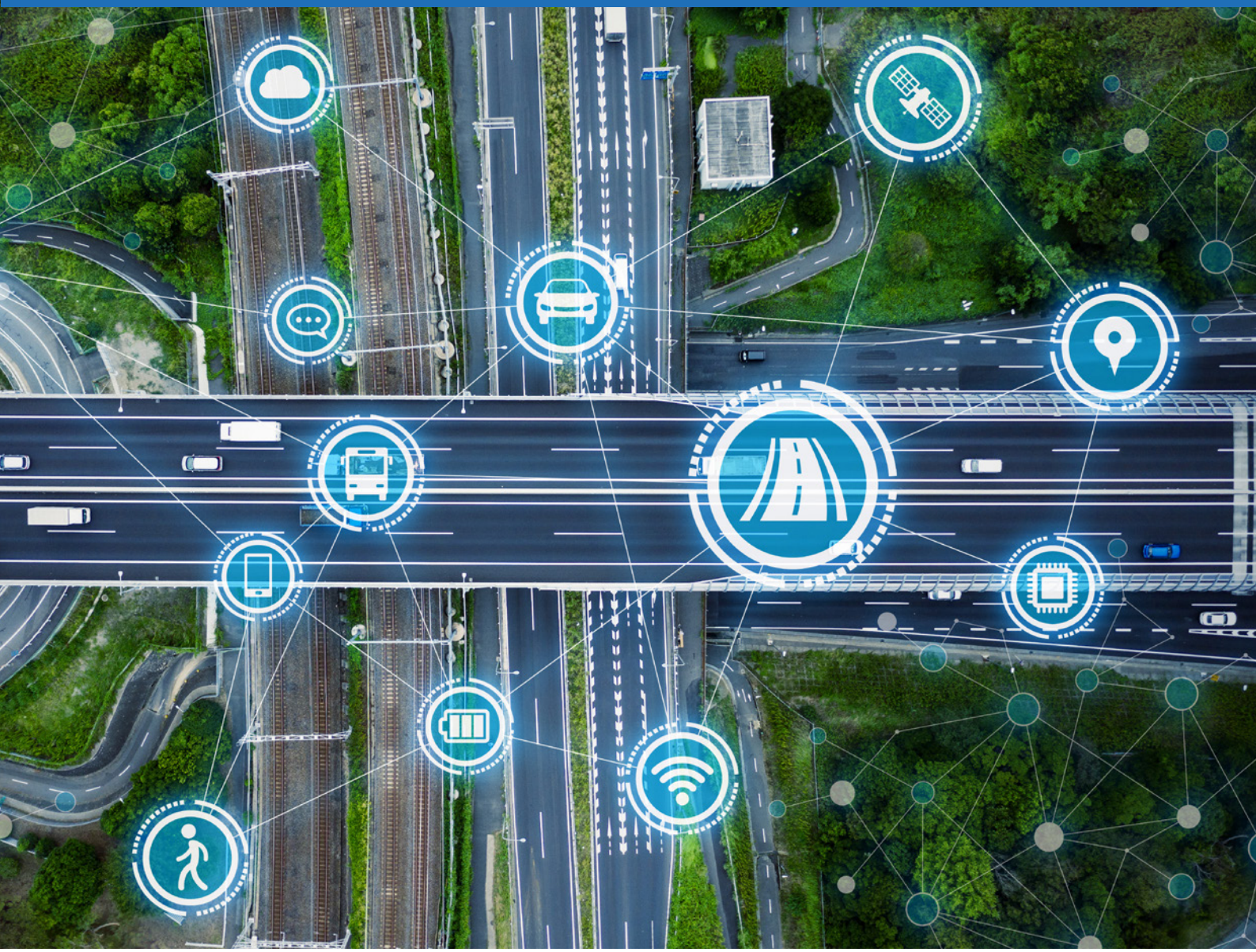
Predicted
Price per m2
Attractivity index 1-10

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!DR74jt
Statistics Norway: https://www.ssb.no/en
Mapping portal: https://kart.ssb.no/

D

2015 projects



Croatia

Croatian Bureau of Statistics

CROATIAN BUREAU OF STATISTICS

Merging statistics and geospatial information in Member States; 2015 project; final report 20 December 2017

KEYWORDS: grid, business register, address

PROBLEM

Having developed geospatial information for population grid-based data from the 2011 census, the Croatian Bureau of Statistics (CBS) noted that users were requesting similar (but unavailable) information from the business register.

OBJECTIVES

The objective of this project was to analyse the (spatial/grid-based) distribution of businesses (legal units) according to their economic activity and legal form.

METHOD

The State Geodetic Administration (National Mapping Agency) was the source of information for maps, centroids and shapefiles. The spatial statistical register (PSR) and the statistical business register (eSPRi) were used to link businesses. Two shapefiles were used, one containing information on 1.6 million addresses, each with an ID and detailed address information, and the other with a 1 km^2 grid.

The PSR contains data on spatial units for statistics at NUTS levels 1 to 3 and for local administrative units (LAU) level 2 (municipalities), as well as for local self-government units, settlements, statistical circles, statistical enumeration areas, streets/squares, house numbers, households and others; these data are kept in alphanumeric and graphical formats.

The eSPRi contains information on business entities (legal entities and natural persons), their parts and groupings, which were institutionally and formally involved in production and financial processes in the national economy. The eSPRi contains the following attributes (depending on the type of unit): a unique identifier, name, spatial features, size, economic activity, number of employed persons and turnover.

Street names taken from source registers (such as the court register) were matched with data from the PSR. Street codes were then taken over from the SPR and saved in the eSPRi. When a street name in eSPRi could not be automatically linked with the official street name in the PSR, the linking was done manually using an application for street coding and the new links were saved in a thesaurus (which was a component in the application). Each new entry in the thesaurus increased the share of automatically coded streets when new addresses were entered into the eSPRi.

Once the addresses of businesses had been corrected and coded in the eSPRi the active legal units in the eSPRi could be linked with geospatial information. Out of 137 thousand active legal units, 120 thousand units were matched using their address information. Thereafter, other matching techniques were used until matching was complete, for example matching streets within grids rather than with their full address, or matching to the centroid of a settlement. Once all of the legal units had been matched, they could be assigned to a grid cell based on their coordinates.

Once a complete dataset was available it was possible to compile indicators for each grid cell, calculating the number of legal units in each cell according to economic activity and/or legal/organisational form. For the latter, six groups were established: companies; co-operatives; associations; institutions; funds and organisations; state government bodies and bodies of local and regional self-government units.

RESULTS

Based on the final dataset a series of maps were developed showing the density of legal units based on 1 km^2 grid cells. These maps showed the density for various activities (such as agriculture, forestry and fisheries; industry and construction; services) and for the six different legal/ organisational forms.

Furthermore, the data are available through the GeoStat application on the statistical office's website.

Figure 1: Business entities in the tertiary sector, situation as of 30 June 2016

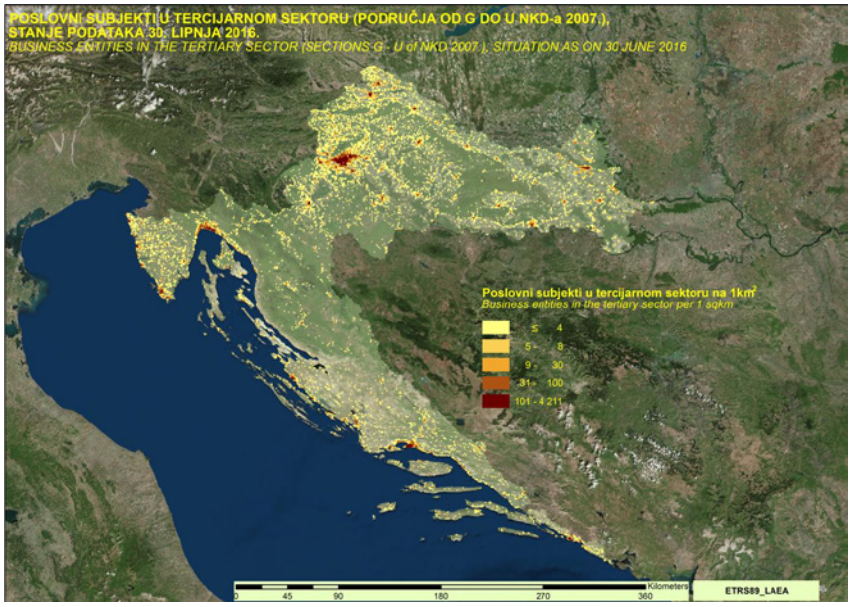
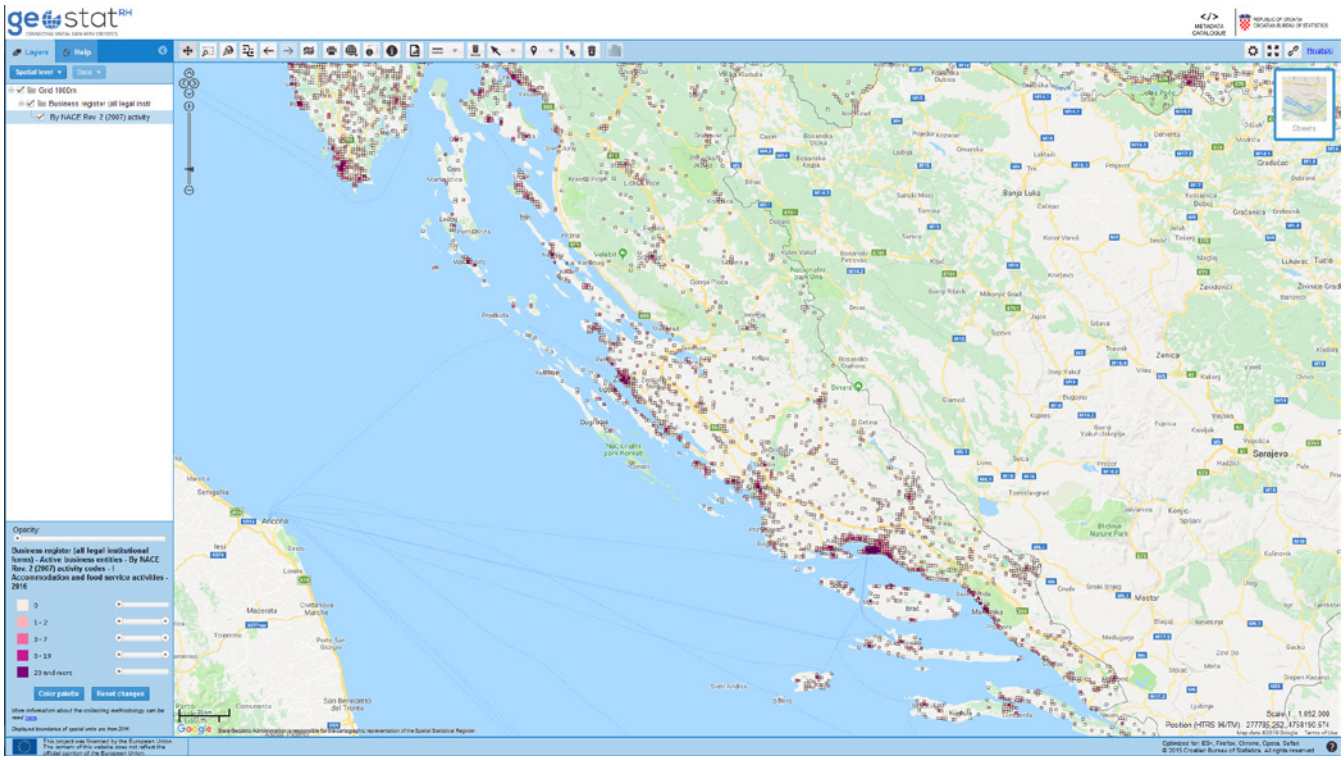


Figure 2: Geostat



FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Rp83RN
Croatian grids register methodology: https://circabc.europa.eu/sd/a/8bf16406-8459-40bb-b5c4-4522e83ce3a9/Methodology%20Croatian%20grids%20Bussines%20register.pdf
Business entities data: https://circabc.europa.eu/sd/a/be6d8c43-dafe-4ef9-8400-8f4bc8a30481/HR.zip
Croatian Bureau of Statistics: https://www.dzs.hr/default_e.htm
Mapping portal: https://geostat.dzs.hr/?lang=en
State Geodetic Administration: https://dgu.gov.hr/

Latvia

Central Statistical Bureau of Latvia

Merging statistics and geospatial information in Member States; 2015 project; final report February 2018

KEYWORDS: address, analytics, demography, OCR

PROBLEM

The Central Statistical Bureau of Latvia (CSB) received a number of requests from researchers asking for a longitudinal data set on population, preferably georeferenced at the level of individuals to enable analyses on migration and depopulation within certain areas of Latvia.

OBJECTIVES

The main objective of this project was the integration of geospatial and statistical information. Specifically, the objective was to geo-reference the results of the 2000 population and housing census (PHC), combine these data with the results from the 2011 census (which were already geo-referenced) and then create a longitudinal dataset suitable for an analysis of developments with respect to migration and depopulation. Internal and external processes required that a continuous (semi-) automated regular update of geospatial data sources should be established.

The second objective of the project was to illustrate how linking geographical and statistical information and corresponding metadata adds value and produces new statistics, in particular for an open data initiative.

METHOD

The 2000 PHC dataset contains information on the place of residence at the level of administrative territories, as well as address information. Optical character recognition (OCR) technology was used to digitise census forms and to prepare a dataset. OCR is known to introduce processing errors in address information (village, street, house name) which need to be eliminated to enable automatic record linking with the state address register information system (SARIS). During this stage, innovative methods were investigated to deal with OCR errors: for example, personal ID numbers were matched between the 2000 PHC and the 2000 population register, and probability was used to assess whether slightly different addresses in the two sources were likely or not to be the same, thereby making it possible to replace addresses with errors in the 2000 PHC dataset using the correct addresses from the 2000 population register. The experience gained through this activity was transferred to internal processes that were implemented for annual geo-referencing of data for the usually resident population and was implemented for the 2016 and 2017 reference years. The addresses in the datasets were then geo-referenced by linking with SARIS. Some problems occurred when addresses in the 2000 PHC dataset had only the village name in the field for the house name and there was a house with that same name: in these cases it was possible that many addresses could be linked to that one house name. A spatial analysis of 2000 and 2011 PHC as part of the quality check helped to identify these and other issues.

For manual linking purposes, alternative data sources were studied (historical data from SARIS, maps from Soviet times) and, where applicable, these were incorporated into an application for manual record linking. The use of an application enabled an operator to use several data sources, including spatial data in order to identify the location of a specific address. A total of 64 thousand addresses were linked manually, 49 thousand were linked exactly, around 5 thousand were linked to a near address, 10 thousand were linked to a village and in 158 cases it was concluded that an address could not be assigned.

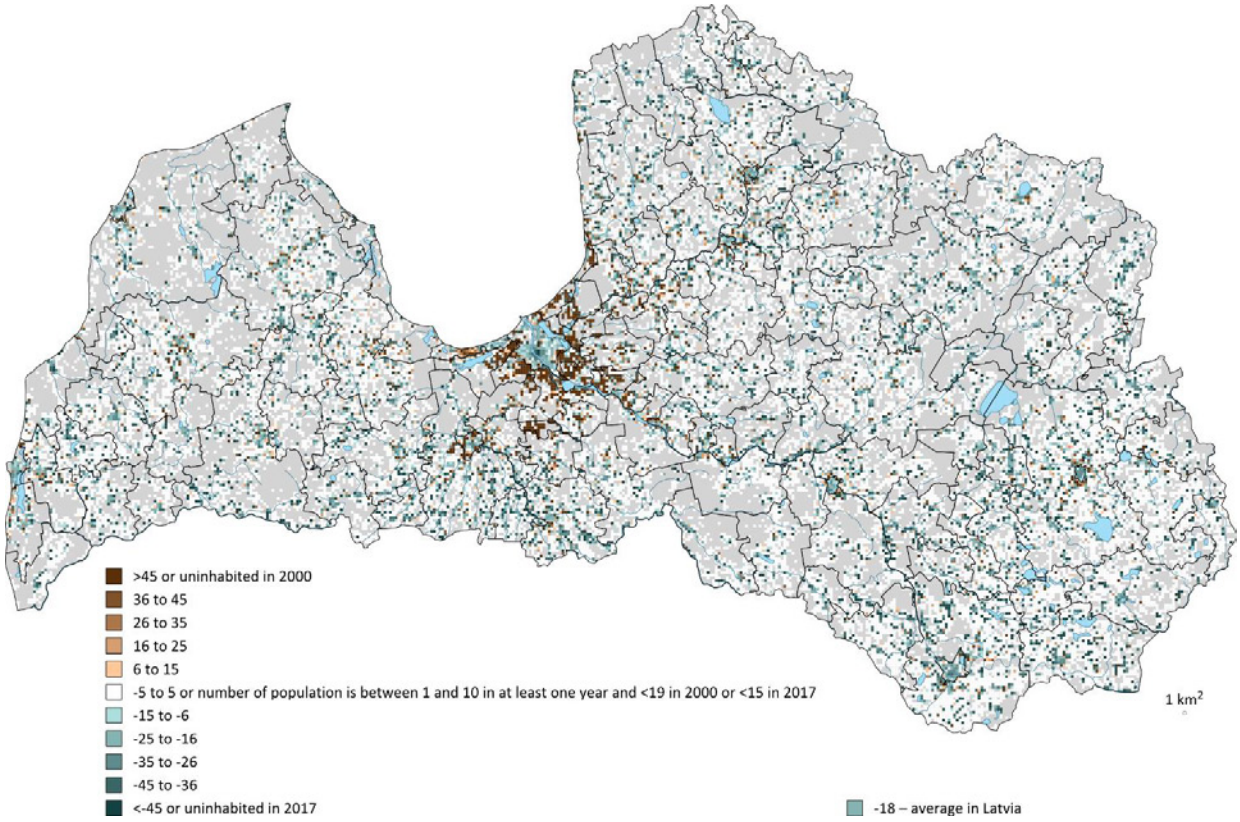
As well as analysing the quality of addresses, an analysis was also made of other personal characteristics, such as working status, education level, knowledge of languages and housing conditions. Some of this information was poorly coded and could be linked to errors made by interviewers, while others were related to OCR problems. Many of the problems were identified by an analysis of maps, showing concentrations of outliers in particular areas. Because of data quality issues, it was decided not to publish the data on personal characteristics from the 2000 PHC dataset and only to disseminate variables from the population register.

RESULTS

Based on the geo-referenced dataset, detailed grid data for population indicators from the 2000 and 2011 PHCs were prepared and published along with estimates for 2016 and 2017. The data published included spatial data for a 1 km^2 grid covering the whole of Latvia and a more detailed 100 m × 100 m grid for cities. Longitudinal tabular data for 2000, 2011, 2016 and 2017 were established, covering demographic indicators based on territorial borders as of 1 January 2018. With this, users could draw objective conclusions on migration and population change (since external factors such as border changes were excluded). A dataset for research was also developed.

The geocoded data for 2000 and 2017 were used to calculate local indicators of spatial association (LISA) for spatial autocorrelation detection at a local level and Getis-Ord Gi* statistics for the identification of hot and cold spots. Hot and cold spots for population change were identified for absolute changes using grid data, spatial clusters and outliers. For the whole of the Latvian territory these proved to be less informative than a bipolar choropleth map of relative changes in population, but they gave reasonable information if limited to smaller areas, like cities. During the analysis phase, issues related to data quality — such as processing and/or measurement errors — were discovered. Due to these errors and limits, a variety of different analyses were carried out, for example, a bivariate analysis of LISA to examine correlations across territorial units between the relative change in population and absolute change in average age. LISA and Getis-Ord Gi* statistics were also used to identify processing and measurement errors.

Figure 1: Change in usually resident population, 2000-2017



FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!gj78uW
Annex I: https://circabc.europa.eu/sd/a/affadd1a-f586-4b41-925e-7e9a1ff2af78/Annex1.pdf
Annex II: https://circabc.europa.eu/sd/a/31914241-56fc-430d-b808-adaf977d7b5d/Annex2.zip
Annex III: https://circabc.europa.eu/sd/a/25611f12-fe08-4ba4-a5c7-deb2740140d8/Annex3.zip
Annex IV: https://circabc.europa.eu/sd/a/2ef6aef1-36a1-4b0d-91b8-23b71bbc7afb/Annex4.pdf
ONS: http://www.csb.gov.lv
Statistical office's maps: https://www.csb.gov.lv/en/statistics/maps-and-spatial-data
Latvian open data portal: http://www.data.gov.lv

The Netherlands

Statistics Netherlands in cooperation with the Dutch Cadastre (1), PDOK (2) and Geonovum (3)

Impact analysis for a table joining service; 2015 project; final report September 2016

KEYWORDS: table joining service, impact analysis

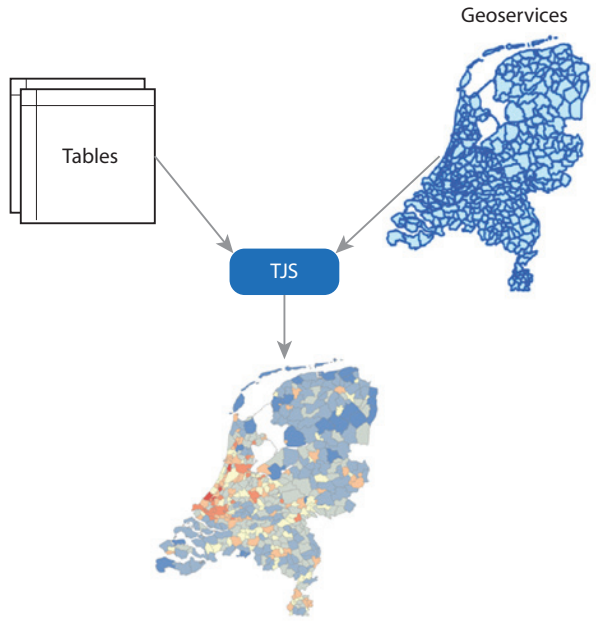
PROBLEM

1. Organisations (and parts thereof) responsible for various tabular datasets do not always have access to online map services or geoservices and are therefore not capable of making maps from these tables.
2. The amount of tabular data from which one can make maps is huge in comparison with the need for maps. Creating maps in advance would be very costly and would make them less up-to-date. Therefore, maps produced from these tables need to be produced when required, rather than in advance.
3. Semantic and technical harmonisation is needed in order to create maps across borders for the INSPIRE regulation for the themes covering population distribution and human health. In fact, this has already been done by means of SDMX services provided by Eurostat, but these tables cannot be viewed or downloaded into GIS systems. In order to do this efficiently one needs an online tool to join these tables to statistical units and create geoservices, which is missing at the moment.

OBJECTIVES

The objective of this project was to perform an impact analysis for a table joining service (TJS) in an environment comparable to the national infrastructure adopted by the Netherlands for the INSPIRE project to see if a TJS could solve (some of) the problems identified above.

Figure 1: Table joining service



METHOD

Statistical data and geographical information have become open and machine readable, although (within the Dutch context) an online tool that linked statistical table services to map services was missing. Such a tool should make it possible to create online statistical maps. This task could be facilitated by a table joining service (TJS): a TJS is an online service that links statistical tables to map services, reproducing what would otherwise be done by a geographic information system (GIS) specialist. The output can be a statistical map service or geoservice that can be saved as an image, or imported into online applications or GIS systems. A TJS connects to source data tables and so its output should be up-to-date and reflect the latest data available. As the output map is available on demand, there is no duplication of stored and published data, nor are maps produced that are not needed.

The project was based on machine readable open data from Statistics Netherlands, combined with a map service with open topographic data for statistical units.

(1) Dutch Cadastre, Land Registry and Mapping Agency.
(2) Public Data on the Map: a collaboration of several government departments who want to publish public geographic data.
(3) National Spatial Data Infrastructure (NSDI) executive committee in the Netherlands.

Based on three scenarios, the project started with an inventory of possible functionalities. It then looked at the impact on potential hardware and software choices on geographical and tabular data and metadata, and on the organisations involved in the project (and other interested parties); this included a financial impact assessment.

Scenario 1: extensive scenario with a TJS plugin on GeoServer, fully supported by the GeoServer community and a client application that is fully maintained; this was the scenario which supports most functionalities.

Scenario 2: this scenario was the same as Scenario 1 but without a fully maintained client application.

Scenario 3: this scenario would be an implementation to fit the self-service environment foreseen by the Public Data on the Map (PDOK) consortium, implementing TJS standards within an Open Geospatial Consortium (OGC) TJS call to:

- upload tabular data and receive input about the connection between the tabular data and the geographic data needed for its presentation on a map;
- process data, converting it to the desired format and making a connection with the geographic data;
- create web features and web mapping services (WFS/WMS), making these services available.

RESULTS

A number of conclusions were drawn from the impact assessment, including the following:

In the longer term it might be a good choice to stimulate the GeoServer community to accept the TJS plugin. The TJS specification was relatively new and was not widely used (at the time of the project). This made its implementation more difficult, as customised software was initially required. In addition, the specification for development needed some improvements, for example, to support warnings for mismatches. Finally, it was considered that there might be difficulties in stimulating the use of a TJS.

When implementing a TJS, it is essential to ensure that use is made of some precise standards. For example, even the difference between a lower and upper case character may cause issues when trying to join information. As a result, users may try to link tables to the wrong geometries; this is particularly an issue if administrative units change every year.

Furthermore, looking at the various scenarios, it was considered likely that some functionalities would be hard to implement, for example, the way a TJS should deal with mismatches and n:m relations; at the time of the project, the standard for a TJS did not appear to have options to deal with this issue efficiently.

Under the third scenario, poor performance with respect to developing web feature services for large datasets was also viewed as a barrier to the implementation of a TJS. Under the PDOK scenario, the system had a download limit of 15 000 records (at the time of writing), which meant that not all datasets were suitable for a TJS; the PDOK consortium was working on increasing this limit. The assessment also highlighted that web features and web mapping services from a TJS were temporary and that this could be considered as a weak point (as permanent results are sometimes required).

Finally, the assessments highlighted that geometric information should be available as open data (which is not always the case in Europe). It noted that a common European TJS might be envisaged as a cost effective way to achieve the goals from INSPIRE ([4]).

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!vR43Pm
Geonovum: http://www.geonovum.nl
Statistics Netherlands: http://www.cbs.nl/en-GB/menu/organisatie/default.htm
PDOK: http://www.pdok.nl
Cadastre: http://www.kadaster.nl
Statline: https://opendata.cbs.nl/statline/#/CBS/nl/

([4]) https://inspire.ec.europa.eu/about-inspire/563.

Austria

Statistics Austria

Merging statistics and geospatial information in Member States — grid-based indicators of accessibility of public utility infrastructure; 2015 project; final report October 2017

KEYWORDS: analytics, grid, accessibility

PROBLEM

In this context, accessibility concerns the ease with which a good or a service (retail trade, education, health, security or leisure) may be reached by the resident population. In order to measure accessibility — defined here as the shortest distance between two points — it is necessary to have a set of grid-based data available at the level of individual buildings.

OBJECTIVES

The objective of this project was to model indicators of accessibility for the resident population with respect to public infrastructure and to produce a set of results that were grid-based datasets.

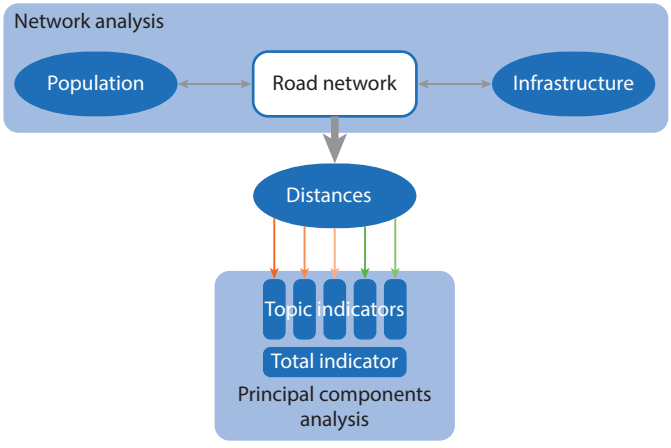
METHOD

The following datasets for infrastructure, demographic indicators and the road network were used.

The main data source for infrastructure was the census of enterprises and their local units of employment (LUE). Units were classified by economic activity and a building ID was also available. For education, school statistics and pre-school statistics were used as they had more detailed information (such as the type of institution). For some special topics, LUE data were cross-checked and if necessary complemented by other data sources with nationwide coverage and high quality, such as healthcare statistics from the Ministry of Health and Women's Affairs and information on pharmacies from the Austrian chamber of pharmacists.

The register of buildings and dwellings (BDR) was the source for building coordinates and building IDs. This register was maintained by Statistics Austria and contained address details of land, buildings and dwellings and held structural data for buildings, dwellings and other usage units including XY coordinates for each address and building. Some coordinates had to be corrected, including, for example, obvious errors such as buildings that were located outside of Austria. A tool was used to match the information from this register with that from other sources; in special cases manual editing was performed. The result was a set of coordinates for all buildings with residents and a set of coordinates for all infrastructure facilities.

Figure 1: Simple workflow



For socio-demographic data, register-based labour market statistics were used. These annual data refer to the situation on 31 October. Data were available for each building concerning the number of resident persons and their socio-demographic characteristics.

For the road network analysis a commercial source (derived from Tele Atlas) was used, including foreign roads in a buffer zone that extended about 50 km around Austria to guarantee the choice of sensible routes including journeys which may transit through neighbouring countries, to complete the information on accessibility for the whole of the Austrian territory, including special cases such as Alpine valleys that may only be reachable from outside of Austria. Some errors in the road network were corrected, for example for some missing minor roads. A speed model developed by Statistics Austria was applied to the road network.

In a second stage, infrastructure facilities were grouped into five broad topics: retailing, education, health, security and leisure. The contents of the datasets were selected and consolidated to one or more data layers allocated to each topic. A total of 90 thousand infrastructure facilities were grouped into 21 layers for the five topics. Then, within a geographic information system (GIS), the quickest route along the road network by private car from a place of residence to the nearest facility was calculated and the time and distance recorded.

The results of this network analysis (distance and time per data layer) were then used as input for principal components analyses along with territorial unit codes (for buildings and three sizes of grid cells), the population by main residence and the assignment of infrastructure data layers to topics. The results were a set of indicators for each topic and each territorial unit as well as an overall (total) indicator for all topics, with indicator values ranging from 0 to 100.

Figure 2: Indicator of accessibility 2014: total for 1km grid

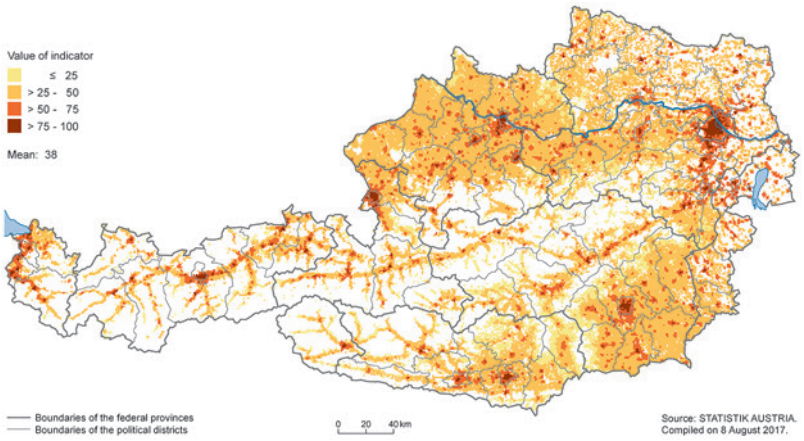
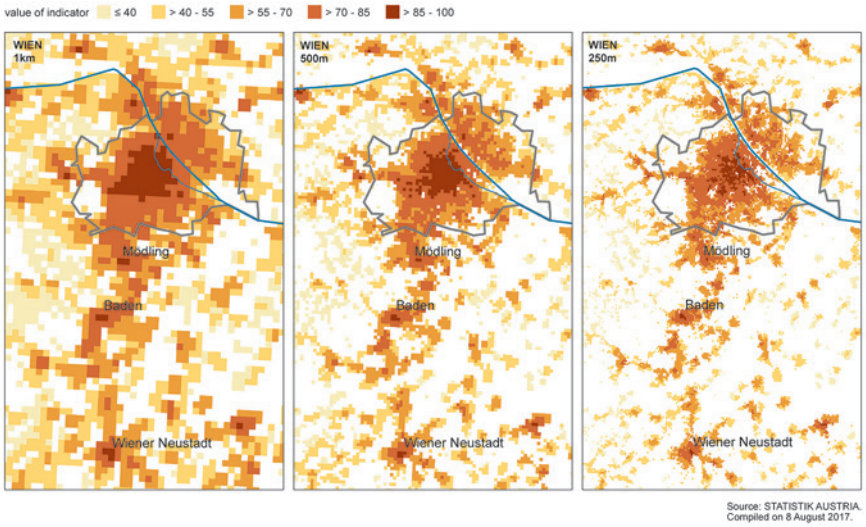


Figure 3: Indicator of accessibility 2014: total for 1 km, 500 m, 250 m grids



RESULTS

The project showed that the calculation of accessibility indicators (total as well as by topic) for public infrastructure at a grid level is possible and produces reasonable results. Indicators by topic showed clearly the differences in the distribution patterns for various types of infrastructure. Comparing the results with territorial classifications such as the urban-rural typology indicated many similarities.

Tests done with smaller grid sizes (500 m*500 m or 250 m*250 m grid cells) showed similar but more distinct results which were particularly interesting for cities.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!Tw47pD
GIS scripts: https://circabc.europa.eu/sd/a/6a58199a-3995-4ca0-b78f-576e6cf52a05/scripts.zip
Erreichbarkeitsindex (accessibility index): https://circabc.europa.eu/sd/a/f26d4be7-0c08-4fa1-8a9a-c53bed14583b/Erreichbarkeitsindex.R
Statistik Austria: http://www.statistik.at/web_en/statistics/index.html
Statistical office's maps — STATatlas: http://www.statistik.at/web_en/publications_services/statatlas/index.html
Regional analysis: http://www.statistik.at/web_en/classifications/regional_breakdown/index.html
Urban-rural analysis: http://www.statistik.at/web_en/classifications/regional_breakdown/urban_rural/index.html

Poland

Statistics Poland

Development of guidelines for publishing statistical data as linked open data; 2015 project; final report 3 January 2018

KEYWORDS: linked open data, URI, local administrative units

PROBLEM

Statistics Poland possesses a vast amount of statistical data that is housed across a range of disparate databases and disseminated using various methods. End users would benefit considerably if this data could be georeferenced and provided in the form of linked open data.

OBJECTIVES

The overall objective of this project was to support the decision-making processes involving the provision of standardised, usable and open geo-referenced statistical data.

Specific objectives:

a) identification of territorial units for which data can be published, including identification of their spatial representation across different years;
b) standardisation of territorial unit identifiers — creating a basis for linking statistical information with geospatial data;
c) feasibility analysis for publishing official statistical resources as linked open data;
d) definition of actions needed to transform existing data to open data formats;
e) description of official statistical resources with metadata in RDF (resource description framework) standards;
f) feasibility analysis for publishing linked open data in the national Geostatistics Portal, including development of guidelines for a linked open data web application.

METHOD

Stocktaking — creation of a data source catalogue

The project involved a stocktaking exercise for various datasets and databases for a range of official statistics analysing these in terms of their content, geo-references and their degree of 'openness'. The stocktaking encompassed all statistical materials published by official statistical entities in whatever form and resulted in a set of metadata, explicitly including information about the use of territorial divisions, data structures and open data.

Territorial divisions

This step involved the identification, harmonisation and generalisation of units for spatial division. The most commonly used divisions were based on the NTS, the Polish equivalent of the NUTS regional classification, and local administrative units (LAUs). Data from the NTS and the geometries for gminas from the National Register of Boundaries (PRG) ([1]) were combined with each other and then with data from the TERYT database in order to produce geometries for regions, voivodeships, subregions and poviats. Geometries for the units composing each of these territorial divisions were imported into a geo-database with harmonised attributes for specific years. A second version of the dataset was developed after generalising (simplifying) shapes in order to reduce file sizes and thus ensure smoother presentation via the internet.

([1]) http://www.geoportal.gov.pl/dane/panstwowy-rejestr-granic.

Table 1: Territorial and statistical unit coding system

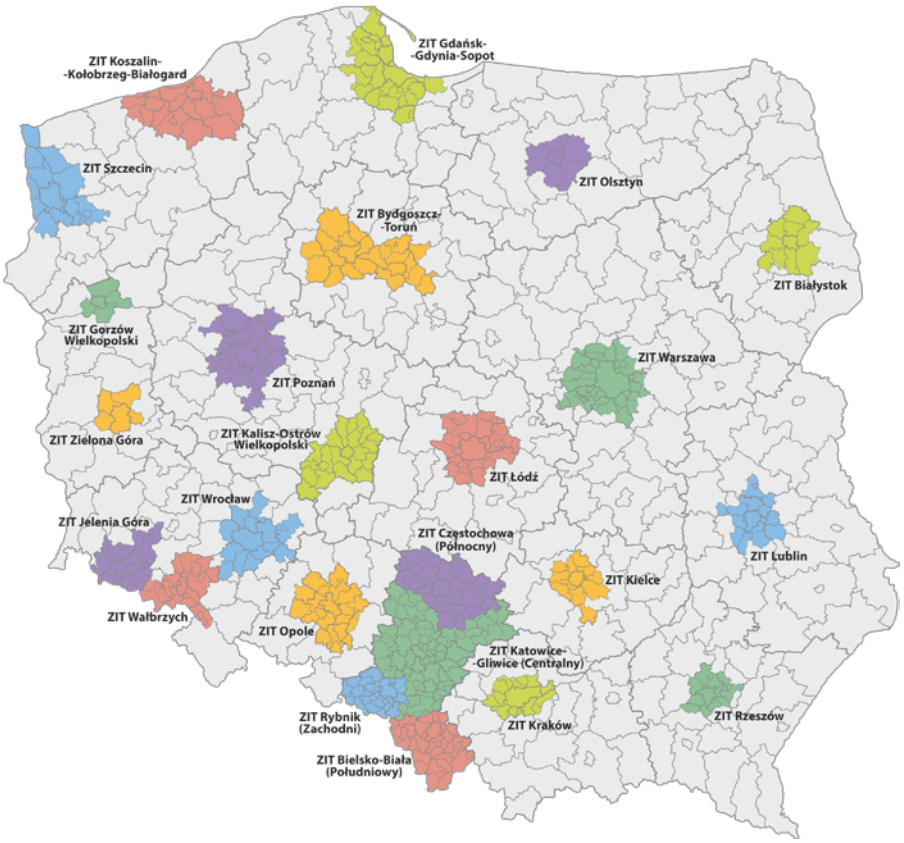
NTS level	NUTS / LAU	Name	Identifier	Unit type	KTS identifier
NTS 1	NUTS 1	Macroregion	1.6	Statistical	**XXXX**0000000000
NTS 2	NUTS 2	Voivodeship (TERYT)	2.6.22	Statistical and administrative	**XXXXXX**00000000
NTS 3	NUTS 3	Subregion	3.6.22.40	Statistical	**XXXXXXXXX**00000
NTS 4	LAU 1	Powiat/cities with powiat rights (TERYT)	4.6.22.40.11	Statistical and administrative	**XXXXXXXXXXX**000
NTS 5	LAU 2	Gmina/part of gmina (TERYT) 1 – urban gmina, 2 – rural gmina, 3 – urban-rural gmina, 4 – urban part of urban-rural gmina, 5 – rural part of urban-rural gmina, 8 – districts of Warsaw, 9 – delegacies of cities: Wrocław, Poznań, Łódź, Kraków.	5.6.22.40.11.01.1	Statistical and administrative	**XXXXXXXXXXXXX**

Concerning other units for territorial divisions, STRATEG (see below) was a tool designed to facilitate the monitoring, development and evaluation of measures taken under cohesion policy. It presented data for non-standard units of territorial divisions, currently providing data for the following functional areas:

- four supraregional strategies;
- two levels of development;
- 20 functional areas related to voivodship development strategies;
- 24 integrated territorial investment areas of functional urban areas (ZIT).

Geometries were prepared for all units for these different territorial divisions.

Figure 1: Integrated Territorial Investment Areas of Functional Urban Areas (ZIT)



Open data technology research

The following conversion process, from data to the development of an ontology to the provision of a service, was developed and researched:

- determine the scope of data publication and methods of searching;
- establish an ontology;
- map the ontology onto existing databases;
- export data to the resource description framework (RDF) format;
- load data to the RDF data store;
- publish data on a linked data server.

Appropriate open source tools and technologies for each of the specific steps of this conversion process were sought after designing the test ontology and then analysed in order to identify the optimum solution for the complete implementation. The following tools were tested:

- the Ontop platform — modelling, mapping and exporting data in the RDF format;
- Apache Jena Fuseki — a SPARQL server;
- Pubby — a linked data front-end (user-friendly interface) for SPARQL end-points;
- OpenCube Toolkit — a set of integrated open source components.

Two proposals for an open data web application for official statistics were developed. The first concerned the development of an open data portal (http://data.stat.gov.pl) for the statistics office using the SPARQL end-point technology (Apache Fuseki) and an extensive user interface. Browsing query results on a webpage interface could be provided by the Pubby software. Data for an RDF data store (Apache Jena) would be prepared using one of the available solutions (Ontop or Python RDFLib) and then imported to the RDF data store. The second option concerned developing the national open data portal (danepubliczne.gov.pl). At the time of writing, the website did not have any solution for SPARQL end-points, but it did provide access to data via an application programming interface (API) ([2]). The Central Statistical Office had a webpage established with hyperlinks to various datasets and services.

The definitions for terms and conditions of use (information on licences) for published data were researched for both options, looking at issues such as the type and scope of licences (for example, public domain or open data licenses) and the compatibility of any licences was assessed when linking to external data sources (for example, the scope of any license, re-licensing conditions, or conditions on the publication of derived work). Equally, certification by the Open Data Institute ([3]) was researched.

Pilot implementation

Three main public statistics databases were reviewed:

- local data bank ([4]) — the biggest set of information on the socioeconomic situation, demography and the state of environment in Poland; it provides access to up-to-date statistical information and enables multidimensional regional and local statistical analysis;
- demography database ([5]) — provides access to statistical information on demography; an integrated data source for the state and structure of the population, vital statistics and migration; enables multidimensional statistical analysis;
- development monitoring system, STRATEG ([6]) — a system designed to facilitate programming and monitoring of development policy; contains a comprehensive set of key measures to monitor the execution of strategies at national, transregional and voivodship level, as well as in the European Union (Europe 2020 strategy); provides access to statistical data for cohesion policy; along with an extensive database, STRATEG offers tools facilitating statistical analysis based on graphs and maps.

From each of these databases a selection of data concerning population by sex and by age was extracted for Poland and its voivodships. Equally, the geometries of Poland and its voivodships were selected for transformation. The data source catalogue developed as part of this project was used as the source of metadata.

([2]) http://apidocs.danepubliczne.gov.pl/.
([3]) https://certificates.theodi.org/.
([4]) https://bdl.stat.gov.pl/BDL/start.
([5]) http://demografia.stat.gov.pl/bazademografia/.
([6]) http://strateg.stat.gov.pl/?lang=en-GB.

A URL structure was established made up of a local server address (for the pilot project) and separate folders for the statistical databases and each of the territorial divisions.

Prior to establishing and implementing an ontology, existing vocabularies were researched, without finding one that could be considered as a reference. As a result, an existing SDMX codelist (CL_SEX) was used for the sex dimension, whereas a new dimension was created for age groups, a codelist was created to specify the population and new dimensions were created for the territorial and statistical units. Namespaces were defined to provide the vocabularies created as part of this project.

After designing the ontologies they were transformed into RDF metadata. Initially, Ontop was used, but this gave unsatisfactory results and so a switch was made to use the Python RDFlib package instead.

Two datasets were created from spatial data, one for Poland as a whole and one for voivodships, both were based on 2016 territorial boundaries. The Python RDFlib package was again used, in this case to convert data from the shapefile format to a coordinate based reference system.

Finally, the three datasets were encoded using the codelists already prepared, with the following structure: **http://[server_name]/[database_name]/2016/POP_SEX_AGE/**. The Uniform Resource Identifier (URI) for each value within a dataset was a combination of the geographic dimension (a 14-digit KTS unit code), sex (based on the SDMX codelist) and age (based on a new codelist). The time dimension was set in the URI at the dataset level. A URI example for the total male population in 2016 for the area of Poland from the local data bank (BDL) is: **http://[server_name]/BDL/2016/POP_SEX_AGE/**10000000000000/Sex-T/TOTAL. In other words, the URI (**in bold**) is followed by a 14-digit KTS code, then by the code for total (from the SDMX CL_SEX codelist) and finally by the code for all age groups (from the Local Data Bank age ontology).

An example of encoding for the total male population in 2016 for Poland from the local data bank (BDL) in Turtle (TTL) format is:

```
<http://10.51.20.122:8090/BDL/2016/POP _ SEX _ AGE/10000000000000/Sex-M/TOTAL> a qb:Observation;
        qb:dataSet <http://10.51.20.122:8090/BDL/2016/POP _ SEX _ AGE>
        sdmx-dimension:age statpl-bdl-age:TOTAL ;
        sdmx-dimension:refArea <http://10.51.20.122:8090/KTS/2016/10000000000000> ;
        sdmx-dimension:refPeriod <http://reference.data.gov.uk/id/gregorian-year/2016> ;
        sdmx-dimension:sex sdmx-code:Sex-M ;
        sdmx-measure:obsValue "18593166"^^xsd:longint.
```

Open data catalogue and dataset metadata

The data source inventory (see above) was transformed into a linked open data catalogue using the DCAT application profile ([7]) for data portals in Europe. The URI was http://[server_name]/KATALOG/.

Equally, each dataset from the data source inventory received a random universally unique identifier (UUID) generated using the uuid Python module. This UUID combined with the URI of the catalogue created the dataset URI.

([7]) https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe.

RDF data store setup and usage

The final stage of the pilot exercise was the creation of a test SPARQL end-point and the creation of a front-end to view (in a web browser) the data store. Apache Jena Fuseki software was used as a SPARQL server. All RDF graphs created within the project were serialised and exported in RDF-XML and Turtle (TTL) formats. All data was loaded as a single Fuseki dataset into the RDF data store using the RDF-XML files.

Finally, a linked open data front-end was set-up using Pubby, which was used to create webpages for each local URI defined in the datasets uploaded to the Apache Jena Fuseki SPARQL server. Each URI created can be viewed with its associated properties as a webpage.

RESULTS

The pilot project provided valuable knowledge on linked open data technologies and vocabularies; several conclusions emerged.

There was no existing reference implementation for statistical linked open data that could be considered as fully appropriate. Existing implementations had some of the following issues:

- lack of integrity between RDF metadata sets published by one authority, probably due to different software or programming components used;
- links to non-existing entities (for example, old ontologies that were not online anymore);
- lack of maintenance (for example, containing data only for a specific reference year).

Figure 2: Example webpage for a dataset

STAT_LOD

Population by Sex and Age Groups

http://10.51.20.122:8090/BDL/2016/POP_SEX_AGE/

Population by sex and age groups - data for voivodships, source: Local Data Bank

qb:DataSet dcat:Dataset

Property	Value
dct:created	2017-10-19 xsd:date
dct:creator	<http://10.51.20.122:8090/metadane/GUS>
dct:description	• Ludność wg płci i grup wieku - dane dla województw, źródło: Bank Danych Lokalnych pl • Population by sex and age groups - data for voivodships, source: Local Data Bank en
dct:modified	2017-12-28 xsd:date
dct:publisher	<http://10.51.20.122:8090/metadane/GUS>
dct:spatial	• kts:Country • kts:Voivodship • <http://publications.europa.eu/resource/authority/country/POL>
dct:temporal	1
dct:theme	<http://eurovoc.europa.eu/385>
dct:title	• Ludność wg płci i grup wieku pl • Population by sex and age groups en
sdmx-attribute:unitMeasure	<http://10.51.20.122:8090/codelist/unit/PER>

At the time of writing, there were no pan-European guidelines for statistical linked open data (for example, which vocabularies or software components to use), although there were several initiatives being run under Eurostat's DIGICOM project.

Some of the tools tested within this project (for example, Ontop or Pubby) became redundant during the course of the project (they were no longer developed), so any implementations based on these might become unstable over time. The Python RDFlib package was considered as sustainable, but it also ceased to be developed during the course of the project.

Linked open data makes most sense if it is connected with as many other data sources as possible. This project used several existing vocabularies and published datasets but a reference statistical linked open data implementation would be a more desired resource.

To achieve this, semantic harmonisation of statistical classifications would be needed. This was considered not only a pan-European issue as it may also impact national data providers, if various datasets have different uses for apparently identical classification elements.

In terms of technology, GeoSPARQL was considered to be an appropriate way to publish spatial data as linked open data. In terms of the temporal aspects, it was considered more complicated.

Separate statistical unit geometries were published for each year, regardless of changes over time. The URIs were constructed based on meaningful identifiers (KTS unit codes). A more appropriate situation may have been to analyse an inventory of statistical unit boundary changes over time and to provide separate geometry instances with non-meaningful identifiers (UUIDs): this would provide a single geometry for a defined period of validity for a unit whose boundary did not change. A prerequisite to do so is information on boundary changes (rather than on the boundaries themselves).

By using existing software and/or programming components it was nearly impossible to produce incorrect RDF metadata files. However, most linked open data producing components allow almost anything to be encoded, so the implementations may not always make sense semantically.

Linked open data implementations based on Python scripts were considered easy to amend, providing flexibility for the future.

RDF vocabulary specifications were considered easier to interpret with a unified modelling language (UML) model. The DCAT-AP specification provided a full UML model of all used classes and properties with a clear indication which of these classes and properties were mandatory, recommended or optional. The RDF Data Cube Vocabulary specification also had a simple graphical representation of some of its classes and their relations.

The pilot project identified that the policy for creating links to the statistical office's information portal needed to be redesigned to allow a linked open data implementation. The landing page and download URLs were based on links to specific webpages (ending in *.html) or links to specific files. To enable a correct linked open data implementation, it was recommended that they should be constructed as URIs.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!RY47cN
Annex I Python scripts: https://circabc.europa.eu/sd/a/28322d20-489d-4a13-93c0-5e30b66f854b/Annex_I_Python_scripts.pdf
Annex II LOD scripts: https://circabc.europa.eu/sd/a/6c51185e-d177-4a87-9e46-e14a38ea0f8f/Annex_I_LOD_scripts.zip
Annex III Python scripts source data: https://circabc.europa.eu/sd/a/8a003114-3fc7-4b10-85dd-3b325d2fbd2b/Annex_II_Python_scripts_source_data.zip
Annex IV LOD graphs RDF XML: https://circabc.europa.eu/sd/a/cc47ec29-841f-4b7d-a9a3-7a123807261f/Annex_III_LOD_graphs_RDF_XML.zip
Annex V LOD graphs TTL: https://circabc.europa.eu/sd/a/b774bc32-2816-411a-975e-9c747857613f/Annex_IV_LOD_graphs_TTL.zip
Statistics Poland: http://stat.gov.pl/en/
Open data portal/mapping portal: http://geo.stat.gov.pl/en/
TERYT register: http://eteryt.stat.gov.pl/eTeryt/english.aspx

Slovenia

Republic of Slovenia
Statistical Office

REPUBLIC OF SLOVENIA
STATISTICAL OFFICE

Merging statistics and geospatial information in Member States; 2015 project; final report 3 January 2018

KEYWORDS: income, health, confidentially, visualisation

PROBLEM

A lack of geospatial data (for example, in the fields of income and health statistics) and the need to develop innovative approaches for treating confidentiality were two problematic issues identified by the Statistical Office of the Republic of Slovenia (SURS). They also identified that a great amount of computer coding would be required in order to develop a fully functioning system for the development of a GIS-based data viewer that allows geospatial statistical data to be published as free open data.

OBJECTIVES

The objectives of this project were to:

- establish two geospatial statistical databases, one on income and the other on health; for the latter develop estimation methods;
- examine and implement innovative approaches to statistical confidentiality;
- upgrade a web application called STAGE, in particular by making it accessible for small touchscreen devices.

METHOD

Geospatial statistical database on income statistics

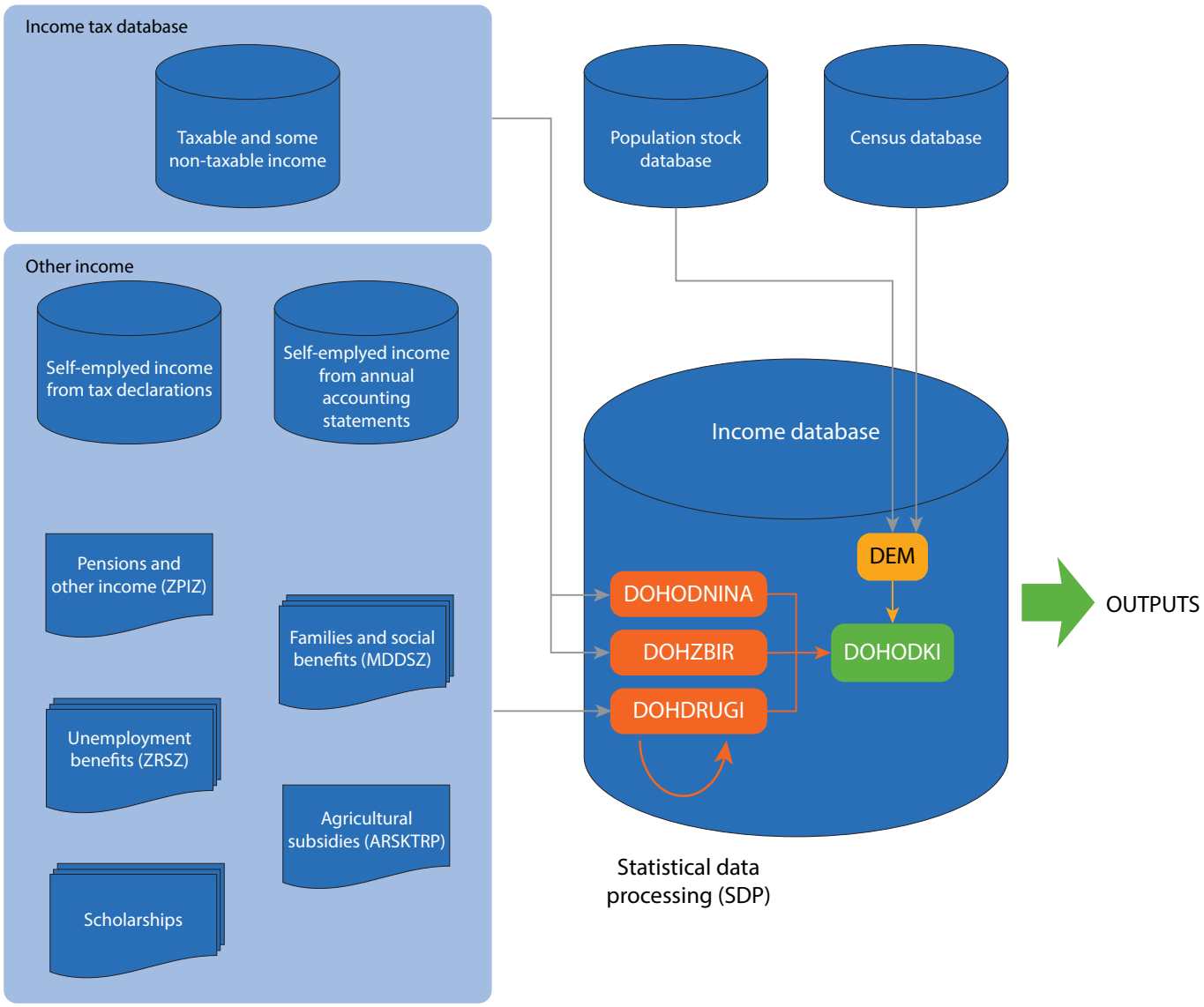
The population stock database and the register-based population and housing census (Census Database) were sources from the statistical office based on the central population register and the household register kept by the Ministry of the Interior (MNZ). It was decided to combine the income database with data from these two sources (for information on 31 December 2014 and/or 1 January 2015) with income data from 2014, available from a variety of sources, including:

- several tax sources from the Financial Administration of the Republic of Slovenia (FURS);
- unemployment benefits from the Employment Service of Slovenia (ZRSZ);
- parental and family receipts, scholarships from the Ministry of Labour, Family, Social Affairs and Equal Opportunities (MDDSZ);
- agricultural subsidies from the Agency for Agricultural Markets and Rural Development (ARSKTRP);
- pensions from the Pension and Disability Insurance Institute of Slovenia (ZPIZ).

An Oracle database was constructed, with procedures to extract, transform and load data from each of these sources.

The population stock database contained demographic data and geospatial data: 24 basic demographic variables (age, gender, residential status as of 31 December and majority of the year), derived variables (education, activity status, profession) and geospatial variables (cohesion region, statistical region, administrative unit, municipality, local community, settlement, spatial district, MID of the house number (XY coordinates)) for the entire population with registered permanent or temporary residence in Slovenia (and also some persons without permanent residence). These data were combined into one table — called DEM — along with the three household variables from the census database (updated every three or four years) and eight derived household variables (which enabled the calculation of poverty indicators).

Figure 1: Income database conceptual model



The DOHODNINA table contained detailed income data from the income tax register. Net income was calculated for each type of income and sums of different kinds of income were also calculated. The DOHZBIR table contained 102 variables from the income tax register on final aggregate income, tax adjustments and reliefs, and income from abroad. The DOHDRUGI table contained detailed income data from all other sources: 195 variables. Data in these three tables were validated and corrected. All records containing income data were linked via personal identification numbers (SID) to the DEM table.

Geospatial statistical database on health statistics

A set of exclusion criteria were established to determine which health indicators would (not) be included in the database. In summary:

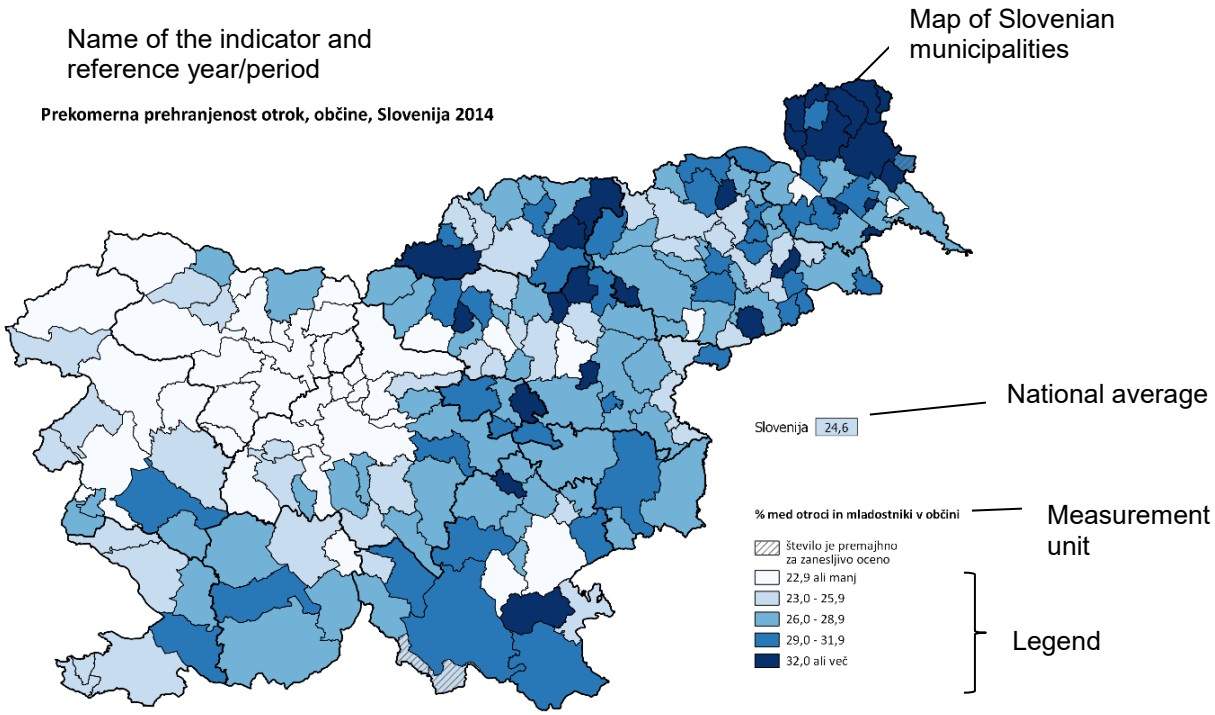
- data were not accessible at the municipality level (LAU 2);
- primary data were considered to be of poor quality;
- data were from a small sample survey;
- there were a small number of phenomena (for example, the low occurrence of specific diseases) in the population;
- data were collected only once, continuous/regular data collection was not expected;
- some similar indicators were duplicated, describing the same area/phenomena;
- some indicators were ambiguous, they did not reflect their purpose sufficiently;
- some indicators did not provide statistical support for decision-making, and would fail to invoke any action.

After discussions with stakeholders, 31 indicators were selected and grouped under four main chapters: risk factors, prevention, health status and mortality. The data sources used for these indicators were scattered among various institutions and among a variety of data environments.

Estimates for the whole set of heath indicators were calculated at the level of municipalities (LAU 2), administrative units (LAU 1), statistical regions (NUTS level 3), and the whole country. For most indicators three- or five-year moving averages were calculated to reduce the variability inherent in the occurrence of rare events, particularly in the smallest municipalities. Furthermore, for some indicators, information by age was standardised (based on the mid-2014 Slovenian age structure). As the data were age-standardised and presented as moving averages, further controls for statistical disclosure were regarded as unnecessary.

Thematic (choropleth) maps were produced for each of the indicators, showing data for 2012 municipalities. These maps were initially made available in a fixed ("printed") format, with plans to disseminate them through the STAGE platform; it was planned to update the database on an annual basis.

Figure 2: Core elements of the thematic map



Downscaling methods for sample surveys

While many health indicators came from exhaustive administrative data, there were a lack of indicators for measuring health behaviour, health determinants and the self-assessment of health and disability. Indicators for these issues generally came from sample surveys, with the drawback that their samples were often too small for an analysis of detailed territorial divisions, such as municipalities. It was considered difficult to simply increase sample sizes due to budget constraints, time constraints and the likely burden on respondents.

Several methodological approaches were available to overcome these challenges, for example, downscaling methods or small-area estimates. Initially, estimates were developed for the following indicators, selected from the 2014 European health interview survey:

- daily smokers of tobacco products;
- self-perceived health status (good and very good);
- binge drinking;
- people living with available help from their neighbours (if necessary).

Auxiliary variables from national healthcare administrative databases (NIJZ) and from the statistical office were used to improve estimates at the municipality level. Some were variables included in a new geospatial statistical database on health statistics. Others were from the statistical office and mainly concerned indicators at the municipality level that were generally associated with health outcomes and behaviours, including for the labour market, demographics, education, income, transport and environmental indicators.

Two modelling strategies were employed during the preparation of small-area estimates: generalised linear mixed models (using municipalities within administrative units as two random factors) and Bayesian modelling using R-INLA (using municipalities with their neighbours and Besag/BYM model), both within a binomial family. Given a choice of around 60 auxiliary variables, research was undertaken to identify which of these variables were associated with the dependent indicator. Several measures of a quality of fit were observed. At first, a dependent indicator (survey data) was modelled with each possible independent variable; then, a model was prepared containing expert-endorsed variables and a few other indicators that had a significant effect or a good fit in the univariate models. Serious deviances or unexpected values in some municipalities were identified and the model was adjusted. Both models (GLMM and Bayesian) were adopted and the value of their estimates were compared.

Statistical confidentiality

Income data are, in general, deemed to be sensitive, and there may be an increased risk of disclosure of income data when combined with geospatial information, particularly at more detailed levels. The statistical office has traditionally published data for NUTS regions, LAUs and also square grid cells with sizes ranging from 5 km^2 to 100 m^2. Traditionally, disclosure has been controlled through the use of cell suppression: if the frequency of observations in a specific cell was below a certain threshold, then the value for that cell was not shown.

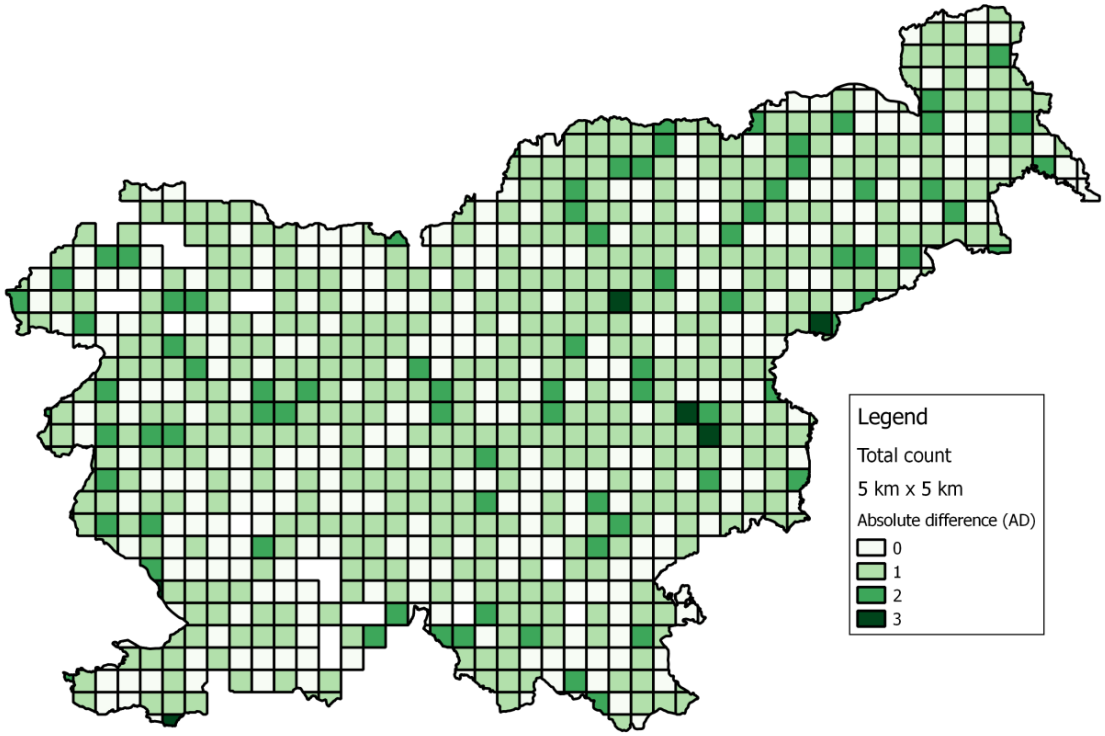
The first step was to review the statistical disclosure control methods most commonly used with microdata and to participate in an international conference on privacy in statistical databases.

Two perturbative methods of statistical disclosure control were selected for testing which were designed specifically with geospatial data in mind. The first method was a record swapping procedure, intended to identify rare individuals at risk of disclosure according to a pre-specified set of variables. The disclosure risk was assessed at each geographical level. Those households where individuals were considered to be at risk of disclosure were then moved geographically by swapping their geographical variables (while keeping their other characteristics unchanged).

The second method chosen for further tests was a perturbative post-tabular method called the cell-key method. This method generated perturbations for small-valued cells in a consistent way, so that the cells consisting of the same individuals were always perturbed by the same value, even in different tabulations.

Preliminary tests were done on a set of demographic variables from the 2011 census. While swapping records only changed the total population count in a small number of grid cells the changes observed were sometimes large, for example if large institutional households were swapped. The cell-key method perturbed most of the original values, but the perturbations conformed to a pre-specified distribution and were mostly small, such that the total count was unchanged.

Figure 3: The spatial distribution of absolute differences between the original and perturbed population count



Further tests for the cell-key method were carried out for income data, looking at income quartiles for total gross income; the geospatial distribution of quartile sizes on a 5 km^2 grid was indistinguishable from the original distribution. The cell-key method, however, was originally intended for analysing frequency data and, it was considered that more generally, the statistical disclosure control method should be chosen to be congruent with the type of statistic it was meant to protect.

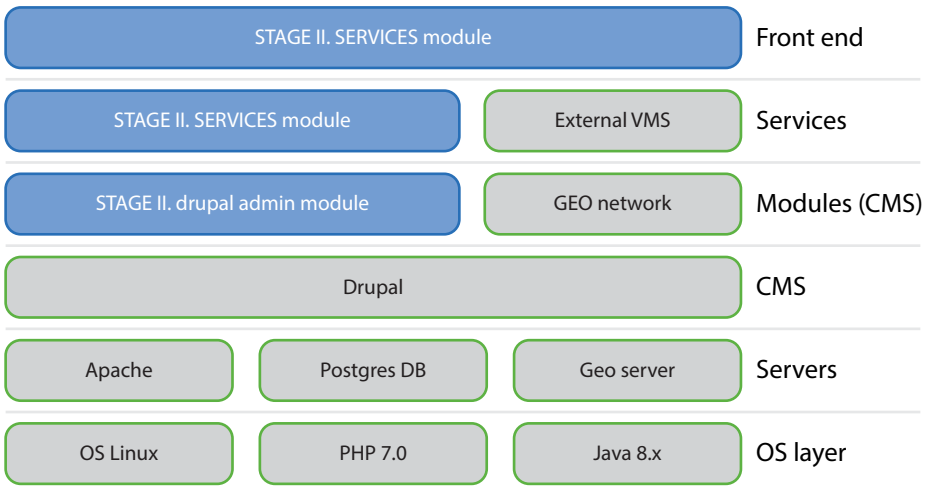
Income data characteristics were explored in more detail and a set of indicators from 2014 were chosen for mapping. The total gross annual income was chosen as the primary variable to be analysed. It was decided that the median (rather than the mean) should be reported as an indicator of central tendency, excluding those persons with no income. Other indicators were: the proportion of income recipients in each municipality within income quintiles(at a national level); the cut-points for quintiles calculated for each municipality; the proportion of people receiving different kinds of income (from employment, pensions and benefits); the median value from different types of income. These indicators were calculated for all municipalities and were also calculated for 1 km^2 grids for Ljubljana and Maribor. The geographies of municipalities and 1 km^2 grid cells were not nested which made the treatment of statistical disclosure control more complicated; the study concluded that indicators published at a detailed geographical level should be chosen carefully. Disclosure risk was also more generally explored for the indicators that were chosen to be published at the municipality level. A simple cell suppression method was chosen, as the frequency tables are mostly considered to be at low risk of disclosure.

STAGE II

Preparations for upgrading the STAGE web application involved the production of detailed functional specifications for all software components. Testing was performed to synchronise the views of developers and subscribers (users) during the development phase. The test environment was used to test the suitability of software performance in an exact replica of what was expected to be used in the production environment. The production environment was used for publishing a beta version of STAGE II.

The system was hierarchically arranged in six structurally connected sets. The bottom OS layer was platform independent, meaning that STAGE II could be installed in any environment that supported Java 8; Java was needed to run the GeoServer. At the time of the study, the Postgres database was used with a Postgis add-on. Apache was used as a webserver. CMS Yii, which was used in STAGE I, was replaced by Drupal (as it was considered the latter offered a more comfortable user experience). GEO network was used as a tool to handle INSPIRE contractual obligations. External VMS services were also included to represent Open street maps and other underlying Web Map Service.

Figure 4: System component diagram



STAGE II administrative core functionalities.

- The menu tree section was used to establish a general time-independent codelist of variable names. The central part of the tab was a graphical interface designed to build a tree menu structure of individual variables as it would be displayed to the STAGE client.
- The variables section was used to import data that corresponded to a single variable defined in the menu tree. Each individual row represented a variable defined for a single spatial unit.
- Variable parameters could be set for a single variable, for a single time intersect, or for a single spatial unit. There was an option to set variable parameters for all variables that corresponded to a single menu tree entry.
- Polygons that corresponded to a single spatial unit could possibly vary over time; therefore, STAGE offered an on-the-fly table joining service when the STAGE client requested data for a choropleth map. In terms of database relations, data on geospatial layers were independent from the data on variables.

STAGE II client functionalities.

- The information page for the selected variable was directly linked to the content shown on the map. At any one time, only a single variable could be displayed on one map in one spatial layer in one time cross-section.
- In addition to the variable description, the user could choose to change the colour scheme, transparency and the classification method. Various methods of data sharing were also provided.
- A choropleth map was rendered according to instructions and the dataset stored in STAGE II administrative core. The STAGE II client was designed to offer various display settings, for example, different ways to determine the class boundaries in each map.
- All values could be displayed in a bar chart and some specific elements could be drawn in the chart. Elements could be defined using different methods: drawing points, circles, rectangles or other polygons.

RESULTS

Geospatial statistical database on income statistics

A new database on income statistics was created. As data were only available for one relatively old reference period (2014), the data on income and poverty for lower geospatial units were not published. It is expected to publish these as a time series when more reference periods become available.

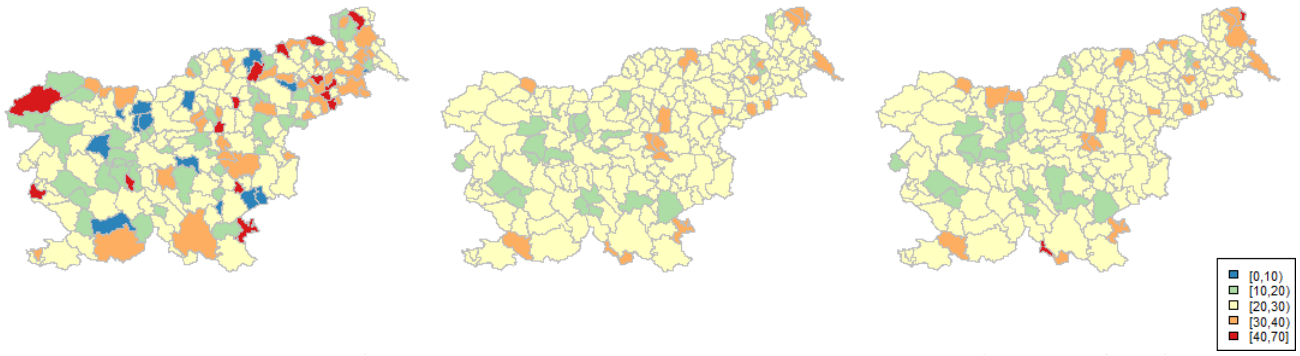
Geospatial statistical database on health statistics

A database on health statistics with 31 indicators at the level of municipalities and local administrative units was made available. It provided an important insight into many public health topics at a detailed geospatial level and the dataset was easy to access for policymakers, thereby promoting a better understanding of the health situation, raising awareness on possible public health issues and planning/promoting activities to improve health.

Downscaling methods for sample surveys

Data models were developed for downscaling methods. For each of the indicators that were modelled, point estimates were considered to be more realistic that simple estimates from survey data, especially for small municipalities. Figure 5 provides an example, showing heterogeneous variety across the country for the survey results and more homogeneous and realistic results from the modelling exercise.

Figure 5: The percentage of current smoking: raw weighted percentages from EHIS (left), GLMM estimates (centre) and R-INLA estimates (right)



Statistical confidentiality

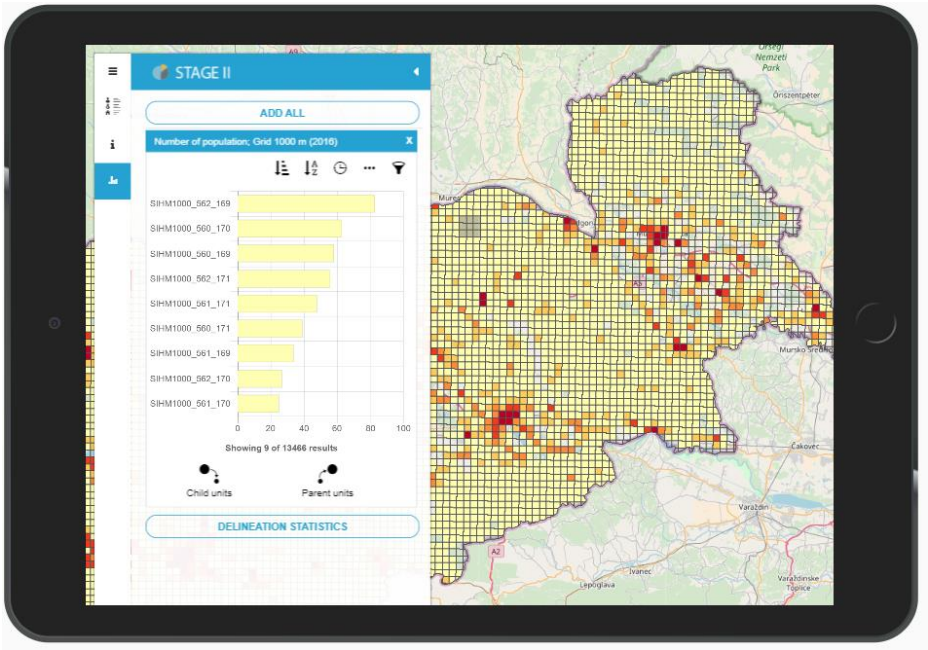
Documentation was prepared for the record swapping method and the cell-key method.

A first set of income indicators were published for municipalities (and grid cells) after disclosure control: the median value for gross annual income; the proportion of people in the first income quintile (at a national level); the value of the first income quintile within individual municipalities; the median value of the three categories of income (from employment, pensions and benefits); the proportion of recipients for each of these three categories of income.

STAGE II

The integrated system for the dissemination of geospatial statistical data — STAGE — was upgraded in cooperation with the Geodetic Institute of Slovenia. All geospatial statistical data included in STAGE were published as free open data. The STAGE source code was made available under the European Union Public Licence.

Figure 6: Stage II client user interface



STAGE II consists of an administrator module and a user interface. The administrator module was, among other functions, intended for: importing statistical and spatial data; hierarchical layouts; naming the displayed variables in a menu; setting visual parameters for impressions to an interactive map. The user interface was developed as a web application, the central part of which was the selected cartographic background and the overlay layer in the form of an interactive map showing statistics for the selected spatial unit. The scan menu placed to the left of the map allowed the selection of variables and dynamic time and space variations, as well as other features. A basic analysis and comparison between the displayed values was also possible.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!fg46dx
Geodetic Institute of Slovenia: http://www.gis.si/en
Republic of Slovenia Statistical Office: https://www.stat.si/StatWeb/en
STAGE (mapping) portal: http://gis.stat.si/
STAGE II (beta) portal: http://gis.stat.si/stage2/#lang=sl
SI-STAT database: http://pxweb.stat.si/pxweb/dialog/statfile1.asp
Statistical data in thematic cartography: https://www.stat.si/TematskaKartografija/Default.aspx?lang=eng
Slovenian statistical regions and municipalities in numbers: https://www.stat.si/obcine/en

Finland

Statistics Finland

Statistics Finland

Statistics on commuting: merging big data and official statistics; 2015 project; final report January 2017

KEYWORDS: analytics, commuting, big data

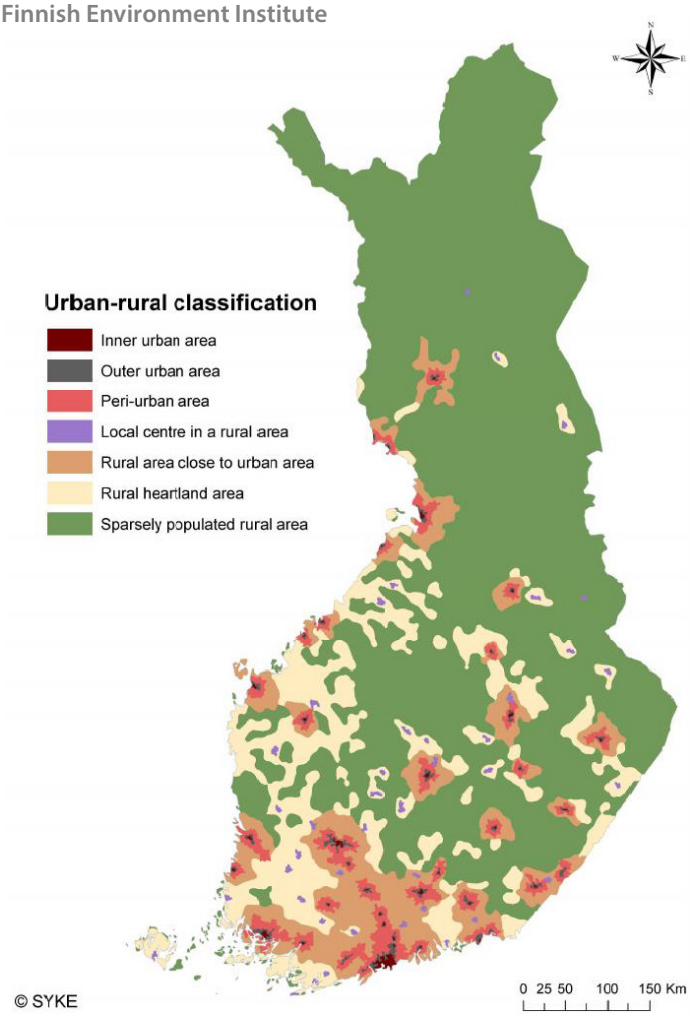
PROBLEM

A lack of official statistics allowing commuting times and distances to be analysed led to the development of a new database that merged big data and official statistics to provide information on commuting patterns and an experimental model for sustainable commuting.

OBJECTIVES

The objective of this project was to conduct a feasibility study for estimating commuting time and distance statistics with traffic sensor implementations: a) comparing a traditional approach with a big data approach for communing time, and b) taking into account public transport to form an annually updated enriched commuting database. The models for the project were built for both commuting time and distance covered by driving, cycling and using public transport within the national network.

Figure 1: The modern urban-rural classification map of the Finnish Environment Institute



METHOD

This study used extensively a set of population statistics stored in Statistics Finland's data warehouse, constructed from several administrative datasets and statistical data files. Dwelling and workplace coordinates were updated annually so that the reference period for statistics was the last week of the year; there was a 22-month lag for workplace coordinates. Home and workplace coordinates were available for 92.6 % of employed residents for the last week of 2013: 2 131 914 people were covered by the framework population out of a possible 2 301 751 in the target population. Commuting estimation statistics presented in the study were limited to those commuters having a maximum journey of no more than 200 km in distance by road between their dwelling and their workplace, giving a total study population of 2 092 132 persons, 90.9 % of the target population.

For the presentation of results, use was made of the Finnish Environment Institute's 2014 urban-rural classification. This was based on 250 m x 250 m grid cells and used seven classes, from sparsely populated rural areas to inner urban areas; these could be grouped into three urban and four rural classes.

Road transport — car and bike

The Finnish Transport Agency (FTA) produces an open national road and street database Digiroad (Digiroad.fi). It contains relatively detailed and accurate data on the location of all roads and streets in Finland, as well as their most important physical features covering a total of 483 thousand km. The dataset is comprehensive in coverage and has a uniform presentation of data; authorities are obliged to update and maintain the data. As the target of the study was the employed population in 2014, September 2014 was selected as the specific target month for these studies as it was outside the principal holiday season. The Digiroad dataset for the fourth quarter of 2014 was used. The FTA also provided data from its systems of automatic traffic measuring devices (hereafter referred to as LAMs), road weather stations, and road weather and surface cameras. The FTA produces official statistics on traffic quantity partly based on the data from LAMs. The target traffic data for this study contained information on average speeds and average numbers of passenger vehicles for each LAM station during a specific time interval of the day, for weekdays between 7 and 8 a.m. over the period from 1 to 26 September 2014.

The speed estimation for using a private car followed a rather complex estimation structure. Speed estimations for each road element were formed by using several feature attributes of the national road and street database (Digiroad.fi), enriched by the FTA's traffic sensor data. Cycle commuting was modelled in a more straightforward, simplified manner, based on the shortest non-hierarchical paths (including separate cycle paths and excluding motorways). The average speed for cycle commuting was assumed to be 17 km/hour and an upper limit of 40 km was set as the maximum distance for commuting by bike.

Public transport

The estimation of public transport accessibility was implemented by utilising two open application programming interfaces. The Journey Planner service covers Helsinki and its surrounding area whereas Journey.fi offers a wide selection of timetables across the whole country for various means of transport, although some timetables were excluded. These services provided information on public transportation timetables for individual users as well as offering large datasets via application programming interfaces. For each person, the coordinates for their home and workplace (as stored at Statistics Finland) were sent to the HTTP GET interface and in return a public transportation route (length, time, vehicle(s) needed) was received in XML format; these were read and saved to a database. The data collection process from each source lasted several weeks. Data were collected for an arrival time of 8 a.m. on 4 January 2016 (for Journey Planner) or 30 May 2016 (for Journey.fi). Journey Planner was used for people who lived and worked in the Helsinki region, while Journey.fi was used for people who lived or worked outside of the Helsinki region.

Walking routes were also considered, especially for short distances, and it was assumed that the speed for walkers averaged 70 m/minute. Note that the time measured was the time taken between departure and arrival points, and did not include any waiting time before departure or after arrival (if earlier than the planned 8 a.m. arrival time.

RESULTS

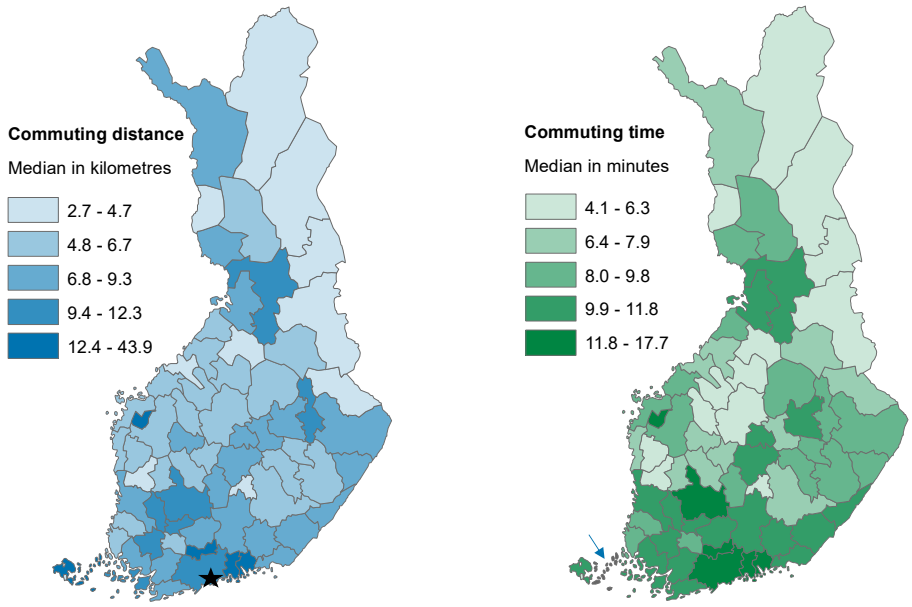
A new database was constructed from the results of this work including the following variables: distance and time for commuting by private vehicle, cycling or public transport across the whole country, and for the Helsinki region public transport commuting distances and times. The commuting distance and commuting time were calculated as point-to-point estimates from almost every employed person's home to a corresponding workplace, with national coverage of approximately 93 %.

Road transport — car

Concerning road travel, the point-to-point linear distance average between the home and workplace was 13.2 km while by road the average was 17.1 km, with an average time of 16.4 minutes. The distribution of the distance travelled was skewed: the median values were 6.0 km linear distance, 8.8 km by road and 11.5 minutes. It was skewed at both ends of the distribution, by a considerable number of short distance commutes, while other commuters sometimes faced lengthy journeys to work. For example, 5.4 % of the framework population had a linear distance between their home and work of less than 100 m.

The Åland Archipelago was a particular case: it has 495 employed persons and their commuting times varied greatly, with some linear routes over 30 km; areas such as this were challenging and required special attention. Other issues occurred when a route could not be found within 300 m of the coordinates of a home or workplace, or where the work or home coordinates were on an island for which there was no registered road or ferry access, or where errors were found in the road network data (such as a dead-end street that was wrongly coded as a one-way street).

Figure 2: Commuting distance and median time

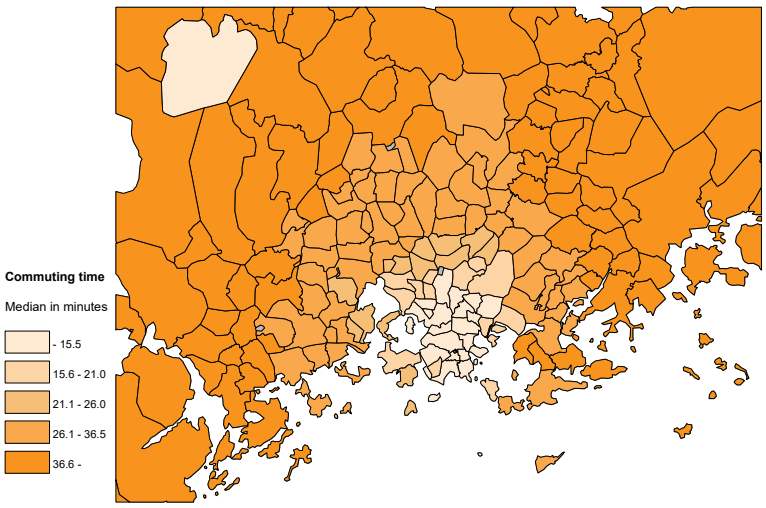


The results were presented in a comparative manner by taking into account the urban-rural classification; they show clear differences by region and area type. Many of the more remote areas were characterised by shorter commuting distances, on average, but higher deviations. People living in inner urban areas generally had shorter commuting distances and shorter times to work except in the largest cities: commuters in the Helsinki region, Turku and Tampere all had longer commuting distances and commuting times, with quite a low level of deviation; conversely, the time and distance covered by commuters from regions surrounding the capital of Helsinki had much higher deviations. Peri-urban areas had longer than average commuting time and commuting distance but with relatively small deviations, while commuters from outer urban areas also had a relatively small deviation in the time and length of their commutes, although these tended to be much shorter than for commuters living in peri-urban areas. The longest median commutes were recorded for commuters living in rural areas that were located close to urban areas. In rural heartlands and in sparsely populated rural areas the time and distance covered by commuters had relatively large deviations: in the countryside many people have their workplace next to (or in) their home, but also many have a registered workplace far away.

Road transport — bike

For cycling, the project calculated that half of the working population could cycle to work in 26 minutes or less. A focus on postal code areas in the capital region (Helsinki, Espoo, Vantaa and Kauniainen, which covered 22.6 % of the employed population across the whole nation) showed that distances and commuting times clearly increased away from city centre locations on the southern coast.

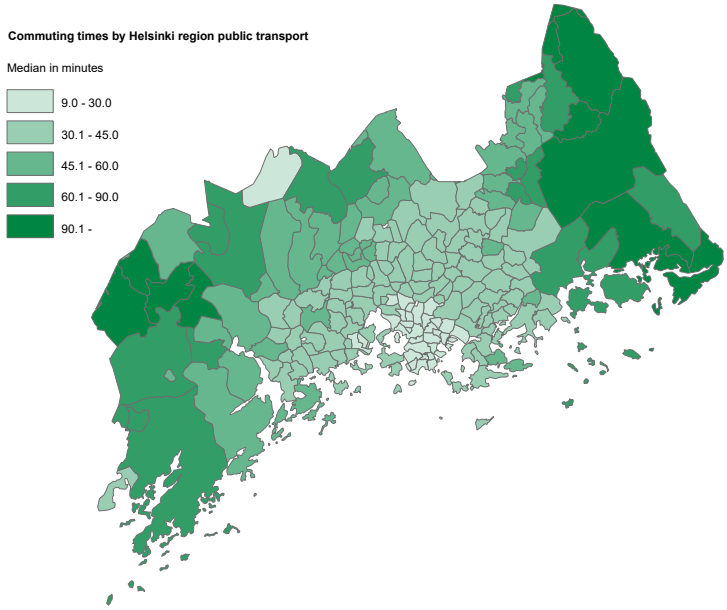
Figure 3: Cycle commuting time by postal code areas in the capital region



Public transport

Journey Planner returned results for 98.8 % of people for whom a pair of coordinates was available, 90.6 % of all employed persons. Journey.fi returned results for 64.9 % of people for whom a pair of coordinates was available, 58.8 % of all employed persons. The results for commuting by public transport were compared with the results for commuting by car with the expectation that they would in all cases be at least as long: in 0.1 % of the cases computed using Journey Planner the commuting route by public transport was at least three times as long (as that for commuting by car) while for Journey.fi this share was 0.9 %. After analysing surprisingly long commuting times and distances for the municipality of Sipoo, these results were excluded as some public transport lines in Sipoo were not available in the journey planning applications and there were errors for some of the transport options; the identification of some errors for this municipality indicated that there could be similar but less obvious errors elsewhere.

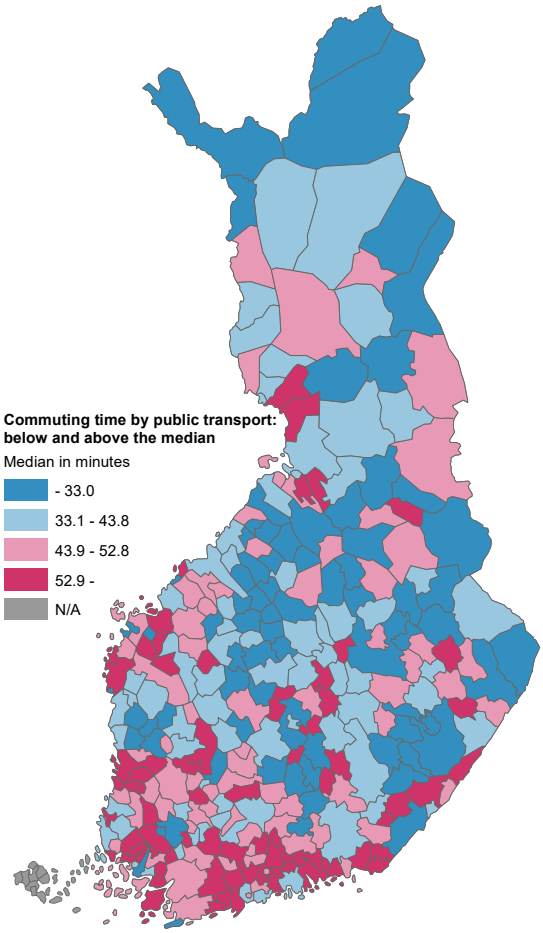
Figure 4: Commuting time by public transport in Helsinki region



The results (particularly for commuting times) were believed to be biased upwards by the inclusion of walking results when no public transport route was found (which in some cases may reflect missing information for available routes), and this was particularly notable for the Åland Islands. Considerably fewer than 1 in 10 of the results from Journey Planner were walking routes, whereas more than one third of the routes from Journey.fi included walking (either a walking route or missing data).

Public transportation distances and times were found to be longer in the Helsinki region. In other parts of the country there was a large variance, suggesting that there was a mixture of a relatively high number of very short commuting distances (to farms and small businesses, for example) and some very long commutes to workplaces; these results were similar to those recorded for commuting by car.

Figure 5: Median commuting time by public transport in Finland municipalities



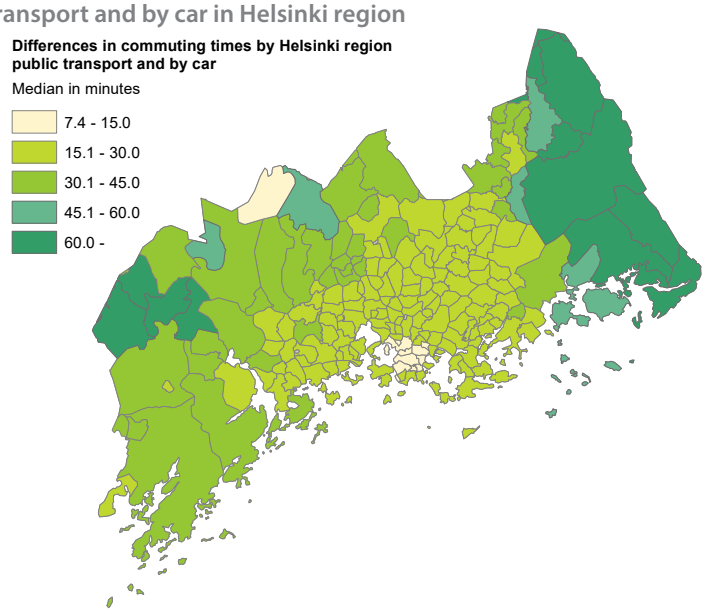
In southern Finland, travelling/commuting times were generally longer than elsewhere in the country. Based on Journey.fi data, the commuting distance from other cities to Helsinki was longer than the median commuting distance to other cities.

Comparison of road and public transport

Comparisons between different transport modes were complicated by a number of factors. The analysis based on public transport might have been hampered by missing information and the strict use of an 8 a.m. arrival time: in case of sparse public transport networks commuters may in reality arrive somewhat earlier or later to take advantage of better routes or (infrequent) connections. Furthermore, the public transport times do not allow extra time for people who arrive early for the first part of their journey. Equally, the analysis by car does not allow for any time to walk to or from the car or any time to find a parking place (which may not be parked directly at home or the workplace), and may always suffer from failing to take sufficient account of congestion and/or traffic controls, particularly in town and city centres during peak travelling times.

An analysis for three transport modes — car, public transport and bicycle — was performed for people living and working in the towns and city covered by the Helsinki Regional Transport Authority. For Helsinki and four of the six other towns in this area, median commuting times were lower for cycling than they were for public transport. In Helsinki and all six towns, cars had the lowest average commuting times, although this may be influenced by the model assuming that commuters by car could travel at average speeds of 20 km/h (which may be too high for such peak travel times, especially in some congested areas).

Figure 6: Median differences of commuting time by public transport and by car in Helsinki region



A second analysis was based on national results: it clearly showed differences in commuting times between the use of public transport and private vehicles for all areas.

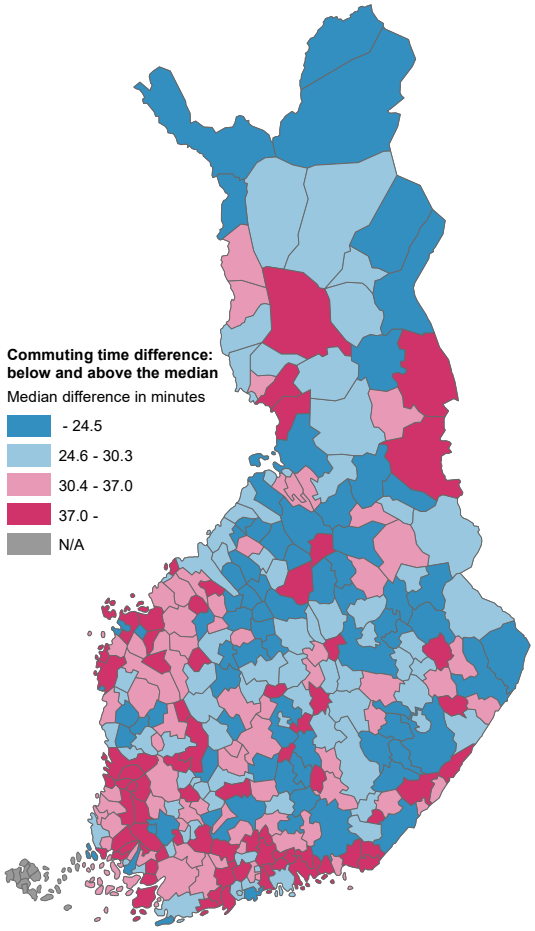
Experimental model for commuting

An experimental model for sustainable commuting was presented as an example of a possible application of this data. This assumed that: i) journeys below 1 km would be made by foot ii) journeys between 1 km and 8.5 km would be made on a bicycle (the upper limit representing 30 minutes of cycling); iii) journeys over 8.5 km would be made using public transport if the journey could be completed in no more than 30 minutes longer than would be required by car and iv) the remainder of the journeys would be made by car. With this model, a median commuting time was estimated at 17.6 minutes, compared with 11.5 minutes if just using a car, with the model estimating that just over one third (36.8 %) of all commuters would use a car, with the remainder using one of the other three transport modes; the relatively high share of car users may reflect, at least to some extent, a potential weakness in the coverage of data for public transport.

FURTHER INFORMATION AND LINKS

Final report: https://europa.eu/!WH78qg
Statistics Finland: https://www.stat.fi/index_en.html
Statistical office's geographic data: https://www.stat.fi/org/avoindata/paikkatietoaineistot_en.html
Statistical office's maps: https://www.stat.fi/tup/vl2010/kartat_en.html

Figure 7: Commuting time difference between public transport and private car, in Finland municipalities



Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by e-mail

Europe Direct is a service that answers your questions about the European Union. You can contact this service

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696 or
- by electronic mail via: https://europa.eu/european-union/contact_en

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU Publications

You can download or order free and priced EU publications at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

Print: KS-FT-19-004-EN-C
PDF: KS-FT-19-004-EN-N

Merging statistics and geospatial information

EXPERIENCES AND OBSERVATIONS FROM NATIONAL STATISTICAL AUTHORITIES, 2012-2015 PROJECTS

Various actions during the course of 2012 — a paper to the European Statistical System Committee (ESSC), a workshop with national statistical institutes (NSIs) and mapping agencies, and a meeting in Prague of Director-Generals of national statistical institutes (DGINS) — resulted in Eurostat deciding to provide a series of grants to statistical authorities to facilitate work on the coordination of statistics and geospatial information.

The association of statistics and geography has the potential to generate information far beyond the simple representation of data on a map. Linking numerical and geo-referenced statistics in spatial analysis may help reveal relationships and phenomena which are difficult to discover by more traditional analyses of statistical databases. *Merging statistics and geospatial information — experiences and observations from the national statistical authorities, 2012-2015 projects* presents details of projects enacted with grants provided during the first four years of this initiative, showcasing the broad range of applications that may be developed using geospatial information.

For more information
https://ec.europa.eu/eurostat/

Publications Office of the European Union

Print: ISBN 978-92-76-03388-2 doi:10.2785/617475
PDF: ISBN 978-92-76-03389-9 doi:10.2785/301125