Creation and use of spatial data by Central Statistical Office of Poland

Mr. Janusz DYGASZEWICZ,

Director of Director of Programming and Coordination of Statistical Survey Department, Central Statistical Office of Poland

Abstract. For several years official statistics in Poland creates and uses spatial data. In order to guarantee the technical conditions required by INSPIRE directive with regard to modern technologies and information standards, and taking into consideration the technological needs of the census and current surveys, the Central Statistical Office had undertaken works to ensure the spatial identification of the included objects on the basis of digital maps and GIS technologies. . Data from databases on statistical address points, statistical regions and census enumeration areas were used to carry out the Agricultural Census 2010 and Census of Population and Housing 2011.In order to visualize the results of Agricultural Census 2010 and Population and Housing Census 2011 on maps Central Statistical Office creates the Goestatistics Portal. For the purposes of spatial visualization, data aggregated to various administrative and statistical units (e.g. municipalities, counties, NUTS regions) was prepared with ensuring statistical data confidentiality.

1. Introduction

For the first time in Poland 28 administrative sources were used in order to obtain the values of the census variables, both at the stage of creating a specification of census units (population and housing census) and for qualitative comparisons. Due to a stable system of identifiers (PIN Personal Identification Number) it was possible to merge data from different registers and from statistical surveys, including sample surveys.

Apart from a wide application of administrative sources, the newest technologies for obtaining data has been used, as a result of which paper forms were completely dropped and replaced by hand-held terminals. Census enumerators equipped with terminals carried out the census on the basis of electronic forms, replacing paper.

Due to the use of digital maps it was possible to manage the census enumerators performing field work with the use of GIS (Geographic Information Systems) tools, where devices with GPS showed on-line on an orthophotomaps the current position of a given enumerator and the address point to which they were to go in order to carry out the census.

Moreover, for the first time the Internet was widely applied in order to allow self-enumeration. It pertained in particular to persons who at a given moment was outside their place of residence or even outside the borders of the country and who was subjected to census obligation. In the case of the Internet self-enumeration, the visit of the enumerator was no longer necessary, which caused a decrease in the costs of the census.

Apart from a direct interview by the enumerator or the Internet self-enumeration, the process of obtaining census data was supported by telephone interviews, with the use of the CATI method.

The whole process was managed by regional centres at the NUTS2, supported by the appropriate managing software based on GIS technologies.

As a result of the use of administrative registers and modern technologies for obtaining data it made possible to reduce the number of enumerators working in the field by over ten times – from approx. 170 thousand in the last census in 2002 to 18 thousand in the 2011 census. This allowed a reduction in census costs by approx. PLN 200 million, i.e. approx. EUR 50 million.

2. Legal basis

In relation to the national population and housing census, the legal Acts currently in force were the following:

- Regulation (EC) of the European Parliament and Council No. 763/2008 of 9 July 2008 on population and housing censuses (EU Official Journal No. L 218 of 13 August 2008)
- The Act on the National Population and Housing Census in 2011 of 4 March 2010 specifying national requirements and the scope of the rights and responsibilities of entities participating in the census.
- The Act of 29 June 1995 on public statistics, where the legal basis and the subject matter of censuses was regulated in a general manner.

3. The implementation model of censuses

Poland, as in most European countries, used the mixed model consisting of merging the data from administrative registers with the data obtained from direct statistical surveys. We claimed that this method was safer and more effective, taking into consideration the present level of development of administrative sources, their quality, and the degree of advancement of methodological work concerning the estimation and imputation of missing data in administrative sources.

Poland, however, was one of the first countries in the world which prepared a totally innovative method consisting of using several of the most modern techniques for collecting census data simultaneously. This pertained mainly to technologies replacing paper with electronic carriers, enabling more effective data collection and abandoning the use of paper in censuses. They included Internet technologies allowing for self-enumeration through the Internet with the use of an electronic form launched both on-line and off-line. A technology made it possible to carry out interviews on-line by phone (the CATI method). Also the enumerators, equipped with appropriate hand-held devices, could use electronic forms without the need to use paper. Hand-held devices made it possible to collect and deliver data on-line and to use digital maps, which eliminated the necessity to use paper maps and situation sketches. A combination of digital maps and aerial photographs with in-built GPS receivers revolutionised the possibilities of preparing and managing the census process before and during the census, and facilitated the carrying out of multi-dimensional spatial analyses with regard to the census results.

All the above-described technologies were at the same time applied and successfully checked during the trial census. Due to the positive results of the trial census, in the actual census in 2011 four channels for obtaining census data were created:

- administrative sources
- the internet – self-enumeration (CAII)
- telephone interviews (CATI)
- a census with the participation of enumerators with the use of hand-held terminals (CAPI)

The last three channels were supported by on-line electronic forms.

4. The use of administrative sources

The following forms of using information systems in censuses were planned:

- Direct data source for surveys,
- Information source for preparing a specification of entities included in the census (an address and housing sampling frame, agricultural farms),
- Additionally, the source of information for:
  - imputation,
  - data estimation,
  - comparisons and determining the quality of data.

The issue of using data from administrative sources required an in-depth recognition of information resources which were found in these sources. An analysis of all the sources and variables potentially useful for the censuses were carried out. The necessary metadata on

approximately 300 administrative registers were collected, of which the 30 most useful ones were selected. For each of these registers separate records were opened, and all variables from these sources were subjected to the utility analysis. The variables were evaluated with regard to their conformity, in terms of definitions and classification, with the dictionaries existing in Polish and EU statistics. Appropriate weights were determined both for the variables and administrative registers from which these variables came from, taking into consideration their utility and quality. The knowledge concerning the quality and utility of variables from different registers was a basis for the rules of merging data, and their estimation and imputation in the operational base of microdata created. The result of this work was invaluable knowledge concerning the utility and possibility of integrating different registers of public administration which the statistical service had at its disposal.

Finally, in the census in Poland we applied 28 sources from Government and Local-Government administration, and from administrators outside public administration such as real estate administrators, housing co-operatives, power distribution plants and telecommunication operators. All the administrators of databases approached the need for statistics related to censuses with understanding and provided access to their information resources for the purposes of the population and housing census in 2011.

5. Address as a universal link of administrative data

One of the basic aims of the national register of the official territorial division of the country (TERYT), kept by the President of the Central Statistical Office, was to ensure an unambiguous identification of territorial entities at various levels of specificity, such as: voivodship, powiat, gmina, city/town, locality, statistical region, census district, street, building, and dwelling.

TERYT enabled the collecting of data for the above-mentioned spatial objects and provided conditions for comparing them and carrying out analyses, which constituted a substantial factor in the implementation of the Directive INSPIRE 2007/2/EC of the European Parliament and Council of 14 March 2007, establishing the structure of spatial information in the European Community (EU Official Journal L 108 of 25.04.2007, pp. 1-13).

In order to guarantee the technical conditions required by this directive with regard to modern technologies and information standards, and taking into consideration the technological needs of the census, the Central Statistical Office had undertaken work in order to ensure the spatial identification of the included objects on the basis of digital maps and GIS technologies. The result

of the modernisation were obtained a numerical description of gmina, powiat and voivodship boundaries (NUTS1, NUTS2, NUTS3, LAU2) and supplementing address identifiers with geographical coordinates x,y of these address points (centroids of buildings).

Introducing x,y coordinates of address points for unit statistical data made it possible to change the previous system of spatial identification of these data and to move from area classification (census districts) to point classification. This had a key, even revolutionary, meaning for the application of geostatistics. Changing the classification allowed a more flexible grouping of data in national censuses for the smallest areas. It also made it possible to create a base of microdata of a spatial nature enabling the carrying out of spatial analyses of various phenomena, concerning, for instance:

- demography (e.g. the average distance between children's and parents' residence, commuting to work, school, distance to hospital etc.),
- urbanisation and planning (e.g. useful in determining the boundaries of urban agglomerations, metropolies, and the drawing up of land development plans),
- agriculture and environment (analysing the structure of crops, environmental pollution),
- the economy (e.g. analysing the effects of burdensome road and industry investments).

Classification of the analyses conducted by points with coordinates x, y also made it possible to become independent of boundaries changes in the regional division of the country, usually resulting in changes in census districts and laborious recalculations. This facilitated a comparative analysis of time series, regardless of the changes taking place in this division. An additional advantage was the possibility of the aggregation of data both in the structure of the NUTS administrative division and the GRID divisions prepared within the GEOSTAT project.

6. Data processing

The process of data processing started from collecting administrative sources from data administrators' registers in the field defined by the appropriate legal Acts. The Polish statistics had the right to access all unit data stored in information sources of the public and commercial sectors. The obtained data included necessary identifiers and personal data supporting the process of merging (linkng) unit data from different sources.

The unit data obtained from registers were converted into statistical registers, simultaneously were subjected to the process of cleaning, de-duplication and standardisation of data. The process was carried out in the DQS SAS environment. At the same time, metadata were collected on the quality of input data obtained from registers, the applied cleaning procedures and the final quality obtained after applying DQS procedures.

The cleaned data were loaded into the Operational Microdata Base as successive logical layers corresponding to the obtained registers.

7. Description of OMB – layers

The basic structure of data in the OBM was a layers corresponding to administrative registers. It was a set of records, each of which related to one census unit (a person, a dwelling, a household). The records included the values of census attributes derived from source data collected from respondents or defined in a different way (e.g. in the process of imputation). Layer could differ between one another with a set of attributes whose values were presented in a given layer. In the first step of processing, before beginning the census, on the basis of source sets and the census frame the layer referred to as a master record was created, consisting of the initial value of the selected subset of census attributes. The values from this layer were transferred to the CAPI, CATI and CAII processes. After completing the census with the use of the CAPI, CATI and CAII processes, on the basis of information collected in the processes proper layer in the database were created. The layer which was saved in the system could serve as the basis for creating new, internal layer in which new attributes (derived from the existing ones) could be added.

8. Procedures and rules of the Golden Record

The Golden Record was a layer from which values, after anonimisation, was transferred for further processing within the Analytical Microdata Base, and was created from many input layers. The census attributes for this layer were selected from various layers already existing in the system on the base on the predefined rules. The selection of the attribute took place for example on the basis of the so-called confidence coefficient applied to the attributes gathered in a given layer. For instance, the confidence coefficient could provide information that, for a given attribute, the value collected in the CAxI process should be utilised first and, if this was not available, the value derived from registers, and if that was not available, the value calculated in the imputation process. The values of the Golden Record could also defined as a result of the imputation process started for this layer.

9. Depersonalisation (personal data safety) and export to the Analytical Microdata Base

Data for all census units were transferred to the output files of the Golden Record of the Operational Microdata Base. Next, each census unit was ascribed a unique and random statistical identifier, which was saved to the output file of the Golden Record and the so-called temporary transition table. The identifier made it possible to restore the collocation of data to the proper census units. The transition table was destroyed after finishing the census.

After that the output file of the Golden Record was exported to the Analytical Microdata Base. Due to the anonimisation (de-personalization) process, the users of the Analytical Microdata Base were able to associate data with census units.

10. Statistical analyses

In this process, on the basis of objective theoretical and statistical knowledge, operations on data sets were carried out for the purpose of summarising and describing data, and exploring data in search of links, structures, systems, factors and concentrations, as well as testing statistical hypotheses and generalising results. With the use of the appropriate information technology tools, various types of data compilations and statistics in the form of tables, statistical measures and graphics were generated. In particular, the following groups of operations were adopted in the process:

- statistical description,
- correlation analysis,
- variation analysis,
- regression analysis,
- model analysis,
- dimension reduction,
- classification,
- non-parametric tests,
- statistical inference,
- spatial analyses.

11. Control census

Following the basic census, a survey was carried out on the basis of a selected sample (approx. 2%) of census entities. Its objective was to check the completeness of the full censuses conducted, the level of reliability of variables obtained from census, and moreover, the impact of various variables (e.g. data-origin source; enumerator, register, respondent) on the quality of obtained data.

12. Census products

Products were defined based on the methodology used to describe so called “hypercubes” introduced by Eurostat. A full list of hypecubes was described in the following documents:

- Regulation (EC) No. 763/2008 of the European Parliament and Council of 9 July 2008 on population and housing censuses,

- Regulation of the Commission (EC) No. 1201/2009 of 30 November 2009 implementing Regulation (EC) No. 763/2008 of the European Parliament and the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns,
- Regulation of the Commission (EC) implementing Regulation (EC) No. 763/2008 of the European Parliament and of the Council on population and housing censuses in the field of a programme of data and metadata deriving from a population and housing census transferred to Eurostat (draft),

Also micro-aggregates, aggregates, tables and "hypercubs" with information required by domestic users were prepared. It is assumed that there were several forms of presenting the products.

13. Publications (forms, types, portals for advanced users)

Dissemination takes place through specialised interfaces, which were developed according to standards required by external users. The following interfaces were distinguished:

- SDMX – according to the standard developed by Eurostat, used mainly for the purposes of integration with the Census Hub,
- Web-Service – used by Local-Government units and economic entities,
- An analytical and reporting tool – includes a set of functionalities making it possible to define reports and analyses in a flexible manner.
- Spatial analyses - The users of the Analytical Microdata Base are able to download metadata on geographical areas according to the chosen area or existing boundary, for example, territorial and statistical division units. The definitions of areas allow the aggregation of data for the dimensions specified using GIS technology

14. Geostatistics Portal

In order to present the results of Agricultural Census 2010 and Census of Population and Housing 2011 Central Statistical Office is preparing a data visualization platform- Geostatistics Portal (GP). Data obtained from the censuses are stored in the micro-Analytical Database (ABM). For the purposes of spatial visualization aggregates are prepared while maintaining statistical confidentiality. Based on the aggregates prepared in ABM Portal will make data available in the form of predefined thematic cartograms. In addition to the spatial analysis prepared in the form of cartogram individual user, within the Geostatistics Portal, will be able to edit their own thematic maps based on any feature of the thematic data and will be able to print maps produced by themselves. Portal will be able to generate spatial analysis areas of given phenomenon based on buffer of the inserted point, line or area, drawn into any polygonal graphics and will give a

possibility to generate summary report based on given query (tabular reports can be supplemented with a map of given area, and charts of selected features).

Geostatistics Portal architecture consist the following subsystems:

- *Geostatistical Database Maintenance Subsystem*, which provides, among other things, access to reference data collected by the Central Statistical Office (National Register of Borders, the National Register of Geographical Names and orthophotomap) as well as gives the possibility of supplying the GeoSTAT database with aggregated data from the ABM
- *System Administration Subsystem*, which provides among other things- users and their access rights administering and georefenced database management;
- *Content Management Subsystem* that enables full functionality for managing and configuring the appearance of the website, publications, articles and statistics, as well as creating the most commonly searched phrases;
- *Publishing and Updating Geostatistical Information Subsystem*, which allows the publication of both thematic and reference map services. Two basic modules of the subsystem are being prepared:
    - *Geostatistical Data Viewer Module (basic module)* that provides the user with predefined statistical cartograms;
    - *Advanced Geostatistical Data Viewer Module* enlargement of the basic module being extended to the possibility of generating questions about any defined area;
- INSPIRE Services Subsystem, which consists:
    - *Metadata-portal* as a separate service to ensure the implementation of statistical branch metadata profile;
    - *Components - OGC Web services* / ISO: WMS, WFS, WCS, CSW.

In addition, in the Geostatistics Portal metadata search capabilities will be implemented by:

- a simple search- by any text;
- advanced search- by using the textual and spatial criteria;
- the range of the data set or service presentation on the map.

Interface prepared for the external user will be primarily *Geostatistical Data Viewer Module (basic module)* as geostatistical data mapping services as a client available at public website. This module in accessible to the user manner presents the phenomenon of pre-defined thematic maps by using cartograms. In addition, there are primers and reference maps containing for example administrative and statistical divisions, and orthophotomaps.

User will have the ability to choose phenomena from a list (prepared in the form of a tree) for which he would like to see its spatial distribution in the form of the thematic cartogram. Furthermore, while selecting phenomena, user will determine the level of aggregation to the selected unit of territorial division (to country, region, vivodship, subregion, powiat or gmina) and select the unit of presented data (eg. by number of farms, animal population or the number of machines). User will have the opportunity to search for themed events by entering a keyword phrase such as "agricultural land." Returns the user to search the full list of events that contain the specified phrase. By selecting events on the generated list, the application will automatically take the user to the correct position on the tree, where it will be possible to further the process of generating cartogram. User will have the opportunity to search for themed events by entering a keyword phrase like "farmland". Data Viewer will return to the user the full list of events that contain the specified phrase. By selecting events from the generated list, the application will automatically take the user to the correct position on the tree, where it will be possible to further the cartogram generating process.

In addition, in the Geostatistics Portal user will be able to:

- use at any time tools to identify any object on the graphical presentation (all objects on all data layers);
- generate a diagram presenting the theme from a lower dimension (the lower position of the tree of thematic events) for the selected object and for the entire cartogram;
- change initially proposed ranges and colors of chosen cartogram;
- print prepared cartogram and to create his own composition of output (to decide whether the information such as scale, graduations and similar should be added or not);
- connect to the Portal other data such as protected areas, roads or rail network through WMS services from external portals such as http://maps.geoportal.gov.pl/webclient/

Advanced Geostatistical Data Viewer Module will have all the functionality available for the basic module but it will be additionally extended by the following possibilities:

- Control access for authorized users from the administrator level;
- Advanced search tools based on attributes or dimensional criteria;
- Metadata search;
- Advanced tools for map features sketching and selecting;
- Length and area measurement tool;
- The presentation of the GIS analysis results;