Merging statistics and geospatial information, 2018 projects - Poland

Statistics Explained

Statistical confidentiality strategy for demographic data; 2018 project; final report 30 December 2020

This article forms part of Eurostat 's statistical report on the *Integration of statistical and geospatial information* .

Problem

The integration of geospatial and statistical information during the statistical production process needs to improve while the spatial distribution of statistics needs to be investigated. This development coincides with the planned introduction of a 1 km x 1 km grid in the 2021 population and housing census.

Objectives

- Develop a uniform and consistent method for data protection in relation to demographic data by creating a statistical confidentiality strategy for data from the 2021 population and housing census, taking into account the geospatial dimension.
- Ensure the quality of the 2021 population and housing census results, including:
  - minimise information loss,
  - maintain consistency of data disseminated via different channels (using the same statistical disclosure control method, avoiding differences between data for the same territorial units),
  - publish data at the most detailed level with respect to statistical confidentiality,
  - preserve information on inhabited/non-inhabited cells of the 1 km x 1 km grid.
- Find a compromise between the acceptable statistical data disclosure risk and usability of the data.

Method

Based on existing reference material, two methods of statistical disclosure control were selected for further analysis within the project: targeted record swapping (pre-tabular method) and random noise cell-key method (post-tabular method). Existing information technology (IT) solutions were tested for these methods.

Two test datasets were prepared for the Podkarpackie voivodship (Subcarpathia province) which was chosen for reasons of diversity.

- The first was a test dataset to simulate the results from the 2021 population and housing census. It was sampled from the latest social survey frame (OBS), dated 31 December 2018, matched with the results from the 2011 census. In order to be included in the test data, a reference address within the OBS was needed which identified an individual dwelling level that was matched with the system for address identification of streets, real estate, buildings and dwellings (NOBC) of Statistics Poland's National Official Register of the Territorial Division of the Country (TERYT). After selection, the test dataset was pseudonymised. The dataset included: total population; sex (males, females); age (under 15, 15 to 64, 65 and over); employed persons; place of birth (in Poland, in another EU Member State, in a non-EU country); and usual residence compared with 12 months earlier (unchanged, both within Poland, moved from outside of Poland).

- The second was a test dataset to be used for developing and testing the statistical disclosure control methods. It was taken from the population records and was also dated 31 December 2018, independently of the OBS. A dataset with the following variables was created: unique statistical number (UNS); X and Y coordinates in Poland's coordinate reference system; sex; age; place of birth (retrieved from the national population register); and usual residence compared with 12 months earlier (calculated by comparing with population records from one year earlier).

- The geocoding of these two test datasets was done with data from:
  - the TERYT register's NOBC subsystem containing i) a dwelling identifier and ii) X and Y coordinates; and
  - the 1 km x 1 km grid spatial dataset in shapefile format containing i) a grid cell identifier and ii) X and Y coordinates of the lower left corner of the grid.

A total of 14 unique hypercubes (cross tabulations) were defined according to the planned specifications for the 2021 population and housing census. All of these concern population data analysed by location and by sex and most also include an analysis by age. The remaining dimensions (such as current activity status, occupation and educational attainment) are included in a smaller number of hypercubes in order to differentiate them. Some of the dimensions, including geographic location, have various levels of detail, referred to as low, medium and high. Data that were required for these hypercubes that were not available in the geocoded test dataset were added from other sources (such as administrative data or older data from the 2011 census).

IT tools to use on the test data were prepared and the parameters of the statistical disclosure control methods adjusted in order to minimise information loss while maintaining an acceptable disclosure risk.

- The test micro dataset was prepared in the required form.

- Computational experiments were conducted on the micro dataset in the R environment for the two statistical disclosure control methods and in SAS (for checking the calculations in R) for the cell-key method.

- The results of three tests – only targeted record swapping, only cell-key method, and targeted record swapping and cell-key method combined – were analysed in terms of risk and utility measures.

Having completed the test on data for the Podkarpackie voivodship, a full-scale test for the whole country was conducted. Again, a variety of data sources were used to prepare the micro dataset, including data from the 2011 census. In terms of processing, one difference compared with the first test was that four rather than three geographical clusters were used.

Results

From the geospatial perspective, the most important aspect of statistical disclosure control on the 2021 population and housing census results is preserving information on inhabited grid cells. By definition, targeted record swapping does not result in information loss for grid cell habitation, whereas the cell-key method modifies the number of persons in grid cells.

- For the test dataset for the Podkarpackie voivodship, in 169 out of 10 417 grid cells (1.6 % of the total) the total number of persons fell to zero due to the statistical disclosure control.

- The application of the cell-key method after targeted record swapping results in a distinct decline in the risk of disclosure, with a small deterioration of the utility measures.

- The significance of information loss depends on the frequency of a specific demographic variable in a given area.

- Calculations on the dataset for the whole country confirmed the conclusions obtained for the test dataset for the Podkarpackie voivodship.

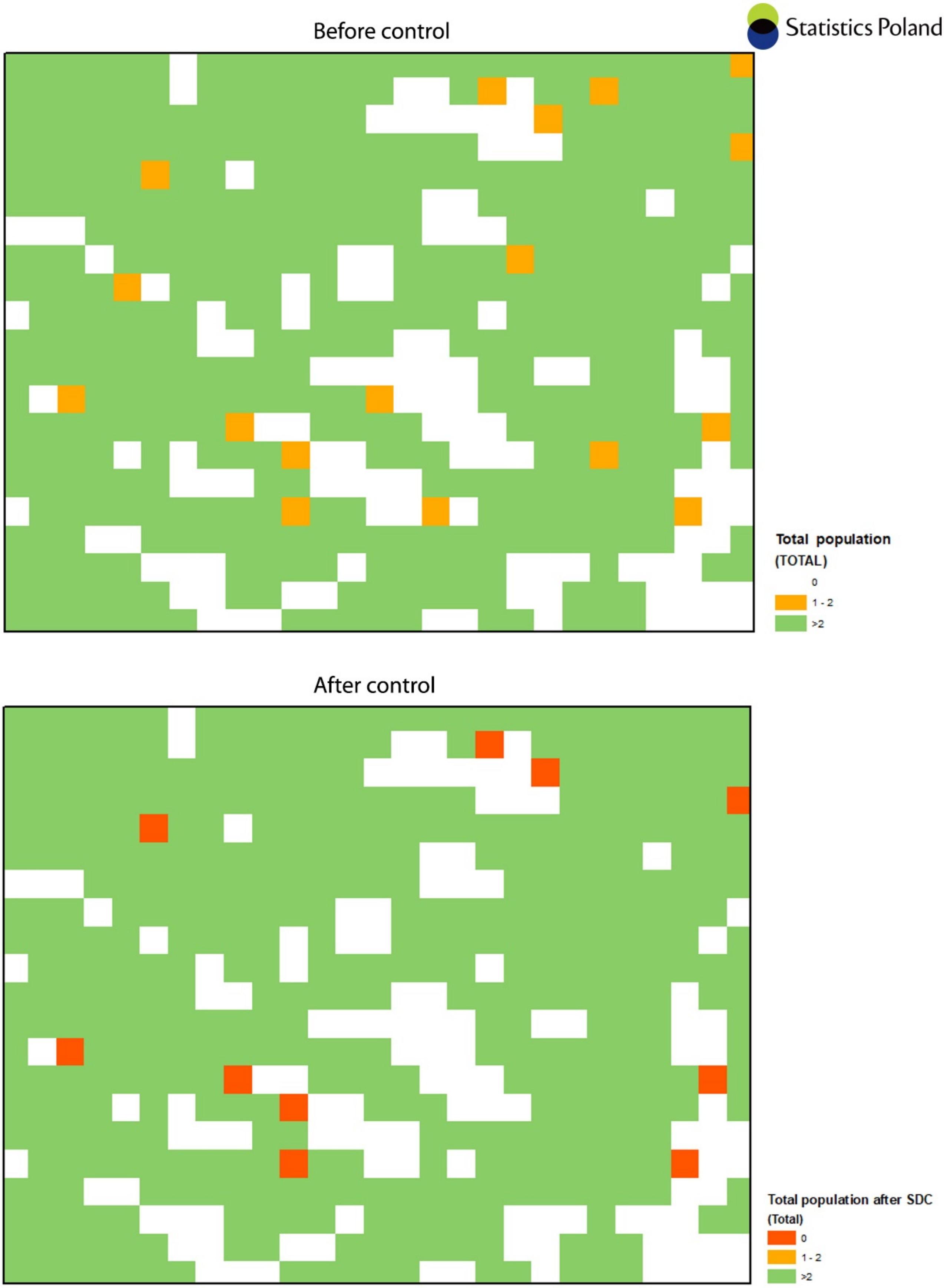

Figure 1: Population count in cells before and after statistical disclosure control, test data, Podkarpackie voivodship

Methodology

- Final report (EU Login required)

External links

- Statistics Poland