STATISTICAL DISCLOSURE CONTROL METHODS FOR HARMONISED PROTECTION OF CENSUS DATA: A GRID CASE

Jaroslav Kraus[1)]

Abstract

The 2011 Population and Housing Census in the Czech Republic was accompanied by a significant change in the technology used to prepare course of the fieldwork, along with changes in how the data are processed and how the outputs are disseminated. Grids are regular polygon networks that divide the territory of country in a grid-like way/pattern into equally large territorial units, to which aggregate statistical data are assigned. The disadvantage of grids is that these are territorially small units that are often minimally populated. This mainly has implications for the protection of individual data, which is associated with statistical disclosure control (SDC).

The research question addressed in this paper is whether data protection (perturbation methods) leads to a change in the characteristics of the file either in terms of statistics of the whole file (i.e. for all grids) or in terms of spatial statistics, which indicate the spatial distribution of the analysed phenomenon. Two possible solutions to the issue of grid data protection are discussed. One comes from the Statistical Office of the European Communities (Eurostat) and the other from Cantabular, which is a product of the Sensible Code Company (SCC) based in Belfast.

According to the Cantabular methodology, one variant was processed, while according to the Eurostat methodology, two variants were calculated, which differ by the parameter settings for maximum noise D and the variance of noise V. The results of the descriptive statistics show a difference in absolute differences when Cantabular and Europstat solutions are compared. In the case of other statistics, the results are fully comparable. This paper is devoted to one specific type of census output. The question is to what extent these results are relevant for other types of census outputs. They differ fundamentally in the number of dimensions (grids have only two dimensions). It would therefore be appropriate to use SDC procedures that allow greater flexibility in defining SDC parameters.

Keywords: population and housing census, statistical disclosure control (SDC), grids

https://doi.org/10.54694/dem.0285 **Demografie,** 2021, **63 (4): 199–215**

INTRODUCTION

Population censuses are a fundamental demographic and statistical task that have long been organised in almost every country in the world. Census programmes are undoubtedly evolving, but the basis remains the same: a census is a survey

1) Czech Statistical Office, contact: jaroslav.kraus@czso.cz.

of the population, houses, flats, and households. The last Population and Housing Census in the Czech Republic took place in 2011 and was conducted in conformity with Regulation No. 763/2008 of the European Parliament and the Council of the European Union. Based on a proposal from the Czech government, the Parliament of the Czech Republic ordered a census by Act No. 296/2009 Coll.

Martin (*Martin*, 2011) evaluated gridded population models using the 2001 Northern Ireland census. He noted that there is a growing interest in the use of gridded population models, which potentially offer the advantages of stability over time and ease of integration with non-population data sources. High-resolution global gridded data for use in population studies were provided in (*Lloyd et al.*, 2017). Recent years have seen substantial growth in openly available satellite and other geospatial data layers, which represents a range of metrics relevant to mapping the global human population at fine spatial scales. Such datasets are vital for measuring the impacts of population growth, monitoring change, and planning policy interventions. (*Lloyd et al.*, 2019) mention the use of global spatio-temporally harmonised datasets to produce high-resolution gridded population distribution datasets. Multi-temporal, globally consistent, high-resolution human population datasets have been used to produce consistent and comparable population distributions to help map sub-national heterogeneities in health, wealth, and resource access, and monitor change in these areas over time. Finally, (*Doxsey-Whitfield et al.*, 2015) took advantage of the improved availability of census data to provide a first picture of the gridded population of the world.

Compared to past censuses, the 2011 Population and Housing Census introduced a relatively significant change in the procedure for preparing the census and in the actual course of the fieldwork, along with changes in how the data were processed and the outputs disseminated. Some methodological approaches have also changed and become more aligned with international recommendations (*CZSO*, 2011; 2013). Although a number of changes have been relatively widely discussed in the literature, one type of output remains somewhat overlooked: census results in a grid network.

In 2012 and 2013, the Czech Republic participated in a project of the European Communities (Eurostat) called Representing Census Data in the European population grid (Geostat). The aim of the project was to create a prototype of the European population grid compiled from national data sets of the results of censuses held around 2010 (in the Czech Republic in 2011) in all participating and cooperating countries and to describe the methodology for generating and displaying these data in the grid.

Three different methods were used to calculate statistical (attribute) data in grids. Because of its high accuracy and the quality of its outputs, the 'aggregation method' is the preferred approach. It is based on the assumption that georeferenced statistical microdata are widely available (provided with X, Y coordinates), with accuracy to the level of buildings and these data are then aggregated within individual grids. In the absence of such spatially localised statistical data, the values for individual squares are derived from the lowest territorial units for which the relevant statistical variables are still available (e.g. municipalities or census tracts); this method is called disaggregation. Finally, if georeferenced microdata are available for only a part of the studied area, then the 'hybrid method' is usually applied, which is based on a combination of the two previously described methods (*Kraus et al.*, 2014).

Grids are regular polygon networks that divide the territory of a country into equally large territorial units, to which aggregate statistical data are assigned (*Klauda*, 2011). In the case of a census, these are squares with an edge of 1 km and aggregations of mean data on the population, although there is nothing to prevent the assignment of data on houses, flats, or households as well.

It is the regular identical shape and thus the identical size of all the cells that is one of the main advantages of grids, which facilitates their mutual comparability in space – for example, across states. Another advantage is long-term stability over time, which contrasts with frequent changes in the definition of administrative units. Networks of squares enable the presentation of statistical data in a very detailed spatial resolution, which brings the advantage of an easier and more accurate analysis of territorial structures (*Kraus et al.*, 2014). However, each method has its advantages

and disadvantages. In the case of grids, it is mainly that they do not coincide with territorial administrative boundaries. This is, of course, solvable, but always only to a certain extent. The second disadvantage is that these are territorially small units, which are often minimally populated, and this is primarily an issue for the protection of individual data - which is associated with statistical disclosure control (SDC).

The research question addressed in this paper is whether data protection (perturbation methods) leads to a change in the characteristics of the file:

- either in terms of the statistics of the whole file (i.e. for all grids), or
- in terms of spatial statistics, which indicate the spatial distribution of the analysed phenomenon.

The issue of SDC is relatively extensive and has been addressed by a number of authors. In this paper, the author often refers to the proceedings of (*Domingo-Ferrer et al.*, 2018), which contain documentation on this issue in relation to the census. A large amount of information, including legal aspects, can be found in (*Hunderpool et al.*, 2012), including calculation procedures for frequency tables. An illustrative way of measuring SDC results, including other useful information, is contained in (*Domingo-Ferrer et al.*, 2006). (*Templ*, 2017) has written a work that is devoted to methods and applications in R in the field of SDC. And (*Thijs et al.*, 2021) have written a practical guide that also deals with applications in R. There is, therefore, sufficient information available for anyone to create own approach to the issue.

Nevertheless, in this paper, two possible solutions to the issue of grid data protection will be discussed. One comes from the Statistical Office of the European Communities (Eurostat) and the other from Cantabular, which is a product of the the Sensible Code Company (SCC) based in Belfast. SensibleCode was involved as a partner in the UK's 2021 population census (*Company*, 2021).

METHODS AND METHODOLOGY

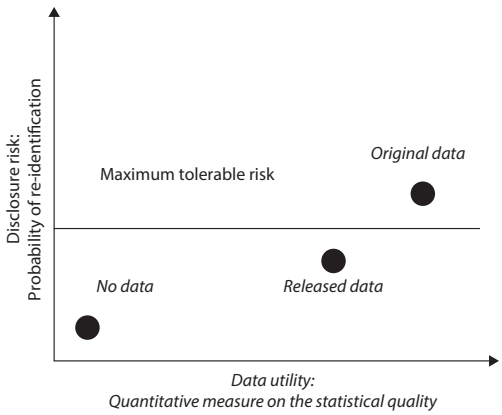
Statistical disclosure control (SDC) is a statistical field that has been developing dynamically in recent years and on which there is already enough good-quality literature. There are many reasons for this development. Disclosure control thinking has to keep up with increases in computing power, developments in matching software, and the proliferation of public and private databases. Statistical offices need to find the right balance between the need to inform society as much as possible, on the one hand, and the need to safeguard the privacy of the respondents on the other (*Hunderpool et al.*, 2012, p. xi). There are several reasons why statistical data protection should be respected. Above all, there are legal regulations that deal directly with the issue of SDC, such as Commission Regulation (EC) No. 831/2002 (*Eurostat*, 2002) of 17 May 2002 implementing Council Regulation (EC) No. 322/97 (*Eurostat*, 1997) on Community Statistics concerning access to confidential data for scientific Purposes, or Commission Regulation (EC) No. 223/2009 of the European Parliament and Council of 11 March 2009 on European statistics. However, there are also a number of other legally relevant documents that focus on this issue (*Domingo-Ferrer et al.*, 2012, pp. 23–35).

There are several ways to address the issue of SDC: traditional methods include tabular data protection or the protection of the output of statistical analyses, and modern methods include microdata protection. This paper is devoted to the latter, and specifically with respect to census output in a grid network. If you work with microdata, i.e. with individual records, then the methods for protecting these data can be divided into several groups. The purpose of all these efforts is to strike a balance between the risk of publishing detailed information and the usefulness of publishing that information.

When assessing SDC methods and their parameters for statistical outputs, an iterative process is carried out. For each method and its parameters, quantitative disclosure risk and information loss measures are calculated. These points can then be plotted on a *Disclosure Risk - Data Utility (R-U) Confidentiality Map*. The optimal SDC method to choose is the one that reduces the disclosure risk to tolerable risk thresholds while ensuring high quality data that are fit for purpose (*Shlomo et al.*, 2006, p. 69).

In the case of microdata, it is possible to define the principles for managing the confidentiality

Figure 1 R-U confidentiality map



Source: *Hunderpool et al.,* 2012, p. 5.

address this issue of confidentiality: Regulation 1588/90 or Regulation 322/97. For statistical disclosure control in the European Union, the following two laws are currently of importance: Commission Regulation (EC) No. 831/2002 and 322/97 on Community Statistics, concerning access to confidential data for scientific purposes.

The purpose of SDC for microdata is to prevent confidential information from being linked to specific respondents when a microdata file is being released. More formally, we can say that, given an original microdata set V, the goal of SDC is to release a protected microdata set **V'** in such a way that:

- the disclosure risk (i.e. the risk that a user or an intruder can use **V'** to determine confidential variables on a specific individual among those in **V**) is low;
- user analyses (regressions, means, etc.) on **V** and **V'** yield the same or at least similar results (*Hunderpool et al.*, 2012, p.23).
- There are two methods to create a protected microdata set **V'**:
- either by masking original data, i.e. generating a modified version **V'** of the original microdata set **V**;
- or by generating synthetic data **V'** that preserve some of the statistical properties of the original data **V**.
- Regarding masking methods, these can in turn be divided into two categories depending on their effect on the original data:
- *Non-perturbative masking:* Non-perturbative methods do not distort data; rather, they produce partial suppressions or reductions of detail in the original data set. Global recording, local suppression and sampling are examples of non-perturbative masking.
- *Perturbative masking:* The microdata set is distorted before publication. In this way, unique combinations of scores in the original data set may disappear and new unique combinations may appear in the perturbed data set; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed data set do not differ significantly from the statistics that would be obtained on the original data set (*Hunderpool et al.*, 2012, p. 33). The whole process of work also depends on whether they are continuous or discontinuous variables.

Random noise is defined by noise probability distributions and by a mechanism to draw from the noise distributions. In its basic form, random noise is generated independently and identically distributed with a mean of zero and a positive variance, which is determined by the statistical agency. A zero mean ensures that no bias is introduced into the original variable. The random noise is then added to the original variable. Adding random noise to a continuous variable will not alter the mean value of the variable for large datasets but will introduce more variance depending on the variance parameter used to generate the noise (*Shlomo*, 2010, p. 3).

Measuring information loss and utility for the SDC decision problem is a more subjective matter. It depends on the users, the purpose of the data, the required statistical analysis, and the type and format of the statistical data. Therefore, it is useful to have a wide range of information loss measures with which to assess the impact of SDC methods on statistical data. These measures include:

- effects on the bias and variance of point estimates and other sufficient statistics,

- distortions to the rankings of variables and univariate and joint distributions between variables,
- changes to model parameters and goodness of fit criteria when carrying out statistical analysis (*Shlomo et al.*, 2006, p. 69).

When assessing SDC methods and their parameters for statistical outputs, an iterative process is carried out. For each method and its parameters, quantitative disclosure risk and information loss measures are calculated. An optimal SDC method is chosen, which reduces the disclosure risk to tolerable risk thresholds, while ensuring high quality data that are fit for purpose (*Shlomo et al.*, 2006, p. 69).

Information loss measures can be classed into two research areas: information loss measures for use by data suppliers so that they can make informed decisions about optimal SDC methods and information loss measures aimed at users so that they can make adjustments to the statistical analysis on modified disclosure controlled statistical data (*Shlomo et al.*, 2006, p. 69).

DATA PROTECTION SOLUTIONS

Eurostat's solution is described in detail in (*Eurostat*, 2017). The relationship to census grid data is also mentioned here. This new geographical variable (e.g. grid id) also needs to be considered from the viewpoint of statistical disclosure control, especially with regard to already existing and used geographical variables. Grid data are particularly useful because they are easy to interpret.

Many grid data will presumably contain zero frequencies. A statistical disclosure control solution cannot alter the spatial distribution of grid data too much. This means that if a few grid cells contain non-zero frequencies in a certain geographical area, they should not be changed very much, and not too many zero grid frequencies should be changed to positive frequencies.

The disclosure risk of statistical data can be quantified using disclosure risk measures. Disclosure risk measures make notions and concepts operational and help to make decisions about the data release. If the disclosure risk is low, a statistical institute might release the data without any change. However, if the disclosure risk is unacceptably high, the statistical institute has to protect the data carefully (*Eurostat*, 2015, chap. 3.1. I, p. 3). The aim is both to protect grids that contain low frequencies of absolute numbers, and to protect low frequencies of attribute values, such as gender, age, marital status, etc. Eurostat's solution is based on the pre-tabular method of targeted record swapping and the post-tabular random noise method. Record swapping is a pre-tabular SDC method, and as such, it is applied to microdata. Some pairs of records are selected in the microdata set. The paired individuals/households are matched on some variables in order to maintain the analytical properties and to minimise the bias of the perturbed microdata set as much as possible. Record swapping exchanges some of the non-equal variable-values between paired individuals/households (*Eurostat*, 2015, chap. 3.1. I, p. 7). The exchanged variables are often geographical variables, and in the case of this paper the grids are used.

Random noise, as a post-tabular method, is defined by noise probability distributions and by a mechanism that draws from the noise distributions. The implementation of random noise as outlined below may involve three 'modules':

- the cell key module,
- the module for determining noise based on cell key and the noise distribution parameter matrix,
- the module to restore additivity (*Eurostat*, 2015, chap. 3.1. I, p. 8).

Cell keys should be drawn from a discrete uniform distribution defined on some integer values (for example, integers between 1 and 100). The process that defines the cell keys has to be consistent, i.e. it must guarantee that the same cell always gets the same key in any hypercube or grid cell or tabulation (*Eurostat*, 2015, chap. 3.1. I, p. 8).

The performance of a random noise method can easily be controlled in a flexible way by means of parameter settings that define the probability distributions. In a typical implementation, the following properties will be required and/or controlled by the parameters:

- noise expectation/unbiasedness property;
- noise variance;
- the property that certain frequencies (e.g. 1s and 2s) should not appear in the perturbed data;

- the property that (structural) zero cells will never be perturbed (*Eurostat*, 2015, chap. 3.1. I, p. 8).

When consistent cell keys are used then the perturbation step leads to consistently perturbed data sets. The ptable files for various settings have been provided by Eurostat for testing. The settings are mainly defined by the maximum perturbation parameter D and by the noise variance parameter V. The ptable provides lists of every combination of cell value and cell key and determines a perturbation value for that cell. The 'p-value' is added to the original cell value, (although most of these changes will be +0) to create the final post perturbation cell value (*Eurostat*, 2015, chap. 3.1. I, pp. 8–9).

Cantabular adds noise to tabular outputs, using the cell-key method, in the same way as the Eurostat methods. Tables are produced dynamically from microdata in real-time in response to a user's query and noise is added deterministically based on a computed cell-key and a perturbation table. Zeros can also be perturbed without affecting any structural zeros found in the data for each query.

The maximum value and variance of perturbation applied are completely configurable via the use of a perturbation table lookup, so different noise distributions can be applied to outputs. In addition to cell-key, Cantabular also includes a disclosure rules language that allows for the real-time checking of table outputs for disclosive cells and the subsequent suppression of outputs per geographic area.

While the Eurostat approach includes a module to restore additivity, Cantabular does not, as this is not possible with a flexible table builder. This loss of additivity can to a small and statistically insignificant degree affect the utility of data for users. This can be avoided by always querying Cantabular for the population counts that are required instead of using Cantabular to create multidimensional hypercubes, which are then themselves queried.

The benefit of taking this approach is that it allows real-time queries for arbitrary cross-tabulations to be made. This is also facilitated by the disclosure rules language, which allows for tables that are still disclosive after the application of cell-key to be automatically suppressed (*Cantabular*, 2021).

INFORMATION LOSS MEASURES

The starting point for measuring the loss of information due to the use of SDC is the evaluation of frequency tables, i.e. the analysis of the differences between the original and the perturbed value. For perturbative methods, we typically measure the maximas, means, medians, and some percentiles of:

- the absolute differences (AD),
- the relative differences (RAD) between original and altered counts in a table, and
- the (squared) differences of the square roots between the original and altered counts.

Counts may be altered because a perturbative protection method has been applied to the data, or because of the effect of cell suppression. The most straightforward way in which to take suppression into account is to impute zeroes for the suppressed count (Eurostat, 2015, chap. 3.1. I, p.10).

According to (*Domingo-Ferrer et al.*, 2006, p. 72), let D^k represent a row (i.e., a distribution) k in a table, and let $D^k(c)$ be the cell frequency c in the row. Let n_r be the number of rows in the comparison. The absolute distance (AD) is then defined as

$$\mathrm{AD(k, c)} := \left| D^k_{pert}(\mathrm{c}) - D^k_{orig}(\mathrm{c}) \right|$$

and the summary statistics per aggregate k mean is defined as

$$\overline{AD(k)} := \frac{\sum_{c \in k} AD(k, c)}{n_k}$$

The relative absolute distance (RAD) is defined as

$$\mathrm{RAD(k, c)} := \frac{\left| D^k_{pert}(c) - D^k_{orig}(c) \right|}{D^k_{orig}(c)}$$

and the summary statistics per aggregate k sum is defined as

$$\sum_{RAD}(k) := \sum_{c \in k} RAD(k, c)$$

Finally, the difference of the square roots is defined as

$$\mathrm{D\_R(k, c)} := \left| \sqrt{D^k_{pert}(\mathrm{c})} - \sqrt{D^k_{orig}(\mathrm{c})} \right|$$

and the suggested summary statistics, e.g. Hellinger's distance (HD), is defined as

$$HD(k) := \frac{1}{\sqrt{2}} \left|\left|D\_R(k)\right|\right|_2 = \sqrt{\frac{1}{2}\sum_{c \in k}(D\_R(k, c))^2}$$

which is used to quantify the similarity between two probability distributions - a namely the original and perturbed datasets. Once these are derived, it is then possible to calculate Hellinger's distance utility (HDU) as

$$1 - \text{HD(orig, perturb)} / \sqrt{(\Sigma \text{orig})}$$

which measures the relative degree of agreement between the original and the perturbed dataset in the interval (0;1).

For both AD and RAD simple descriptive statistics like max, mean, and median, the percentiles p60, p70, p80, p90, p95, and p99 would be calculated. In addition, the cumulative distribution function $\mathbf{F_{RAD}}(\boldsymbol{r})$ proportion of cells with relative absolute difference less than (r) could also be calculated. These measures are based on the idea that if the synthetic and original data are similar, data set membership should be indistinguishable between the two data sets.

Another statistical analysis that is frequently carried out on tabular data are tests for independence between categorical variables that span a table. The test for independence for a two-way table is based on a Pearson Chi-Squared Statistic (*Shlomo*, 2006, p. 214). This statistic defined for i is from 1 to s and the summation for j is from 1 to r, is formulated as

$$Q_p = \sum_{i=1}^{s}\sum_{j=1}^{r}\frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

where

$$m_{ij} = E\{n_{ij} \mid H_0\}\frac{n_{i+}\, n_{+j}}{n}$$

is the expected value of the frequencies in the i th row and j th column.

Measures of association when one or both variables are nominally scaled are more difficult to define, since you cannot think of association in these circumstances as negative or positive in any sense. However, indices of association in the nominal case have been constructed and most are based on mimicking R-squared in some fashion. One such measure is the uncertainty coefficient, and another is the lambda coefficient (*Stokes et al.*, 2012, p. 129).

The asymmetric lambda λ (Columns|Rows) is interpreted as the probable improvement in predicting the column variable Y (perturbed data) given knowledge of the row variable X (original data). The range of the asymmetric lambda is $0 \le \lambda(C|R) \le 1$. The asymmetric lambda (C|R) is computed as

$$\lambda(C|R) = \frac{\sum_i r_i - r}{n - r}$$

and its asymptotic variance is

$$\text{Var}(\lambda(C|R)) = \frac{n - \sum r_i}{(n-r)^3}\left(\sum_i r_i + r - 2\sum_i (r_i \mid l_i = l)\right)$$

The nondirectional lambda (symmetric) is the average of the two asymmetric lambdas, (λ(C | R) and (λ(R | C). Its range is $0 \le \lambda \le 1$. The lambda symmetric is computed as

$$\lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c} = \frac{w - v}{w}$$

and its asymptotic variance is computed as

$$\text{Var}(\lambda) = \frac{1}{w^4}\left(wvy - 2w^2\left(n - \sum_i\sum_j (n_{ij} \mid j = l_i, i = k_j\right) - 2v^2(n - n_{kl})\right)$$

The uncertainty coefficient U is the symmetric version of the two asymmetric uncertainty coefficients. Its range is $0 \le U \le 1$. The uncertainty coefficient is computed as

$$U = 2(H(X) + H(Y) - H(XY)) / (H(X) + H(Y))$$

and its asymptotic variance is

$$\text{Var}(U) = 4\sum_i\sum_{ij}\left(H(XY)\ln\left(\frac{n_i n_j}{n^2}\right) - H(X) + (H(Y))\ln\left(\frac{n_{ij}}{n}\right)\right)^2$$

where $H(X)$, $H(Y)$, and $H(XY)$ are defined in the previous section. See (*SAS Stat*, 2021) for the completed description.

For each measure, the asymptotic standard error (ASE) has been calculated, which is the square root of the asymptotic variance denoted by the variable. If the sample size is adequate, then the measure of association is approximately normally distributed, and the confidence intervals of interest can be calculated as

$$\text{est} \pm z_{\alpha/2} \cdot \text{ASE}$$

where est is the estimate of the measure, zα/2 is the 100 (1-α/2) percentile of the standard normal distribution, and ASE is the asymptotic standard error of the estimate (SAS Stat). In this case, 95% confidence interval was used.

The Gini index (or Gini ratio) is a measure of statistical dispersion and it is the most commonly used measurement of inequality preferably used in economics. It measures the inequality among values of a frequency distribution. An index of zero expresses perfect equality, where all the values are the same, and an index of 1 (or 100%) expresses maximal inequality among the values. The sample Gini coefficient was calculated using the formula:

$$G = \frac{1}{2\bar{X}n(n-1)} \sum_{i=1}^{n} (2i - n - 1)X_i$$

where X_i are the sizes sorted from smallest to largest, X1 ≤ X2 ≤ Xn (*Dixon*, 1987).

FROM GRIDS TO SPATIAL STATISTICS

In the case of grid data, it is also necessary to take into account spatial measures, which measure the degree of spatial distribution both original and perturbed data sets:

- the global autocorrelation rate
- the local autocorrelation rates.

Spatial autocorrelation may be a result of unobserved or hard-to-quantify processes, combined in various places, and together the causing spatial structuring of a given phenomenon. If there is a spatial autocorrelation, it is determined by examining whether the variable value for a given (e.g. geolocalised) observation is associated with values of the same variable for neighbouring observations (*INSEE*, 2018, p. 67). Spatial autocorrelation may be positive or negative or there may be no spatial autocorrelation among the given data. Spatial autocorrelation can be measured globally or locally; both ways assess the same thing – i.e. whether there is a spatial correlation of a given phenomenon – but they are not the same.

There are different ways of measuring spatial autocorrelation; Moran's I is often used. The principle of computation is that it takes into account the difference between the value of the variable and the average of values of that variable for a given area (e.g. neighbourhood). Moran's index is the preferred approach (compared to others), because it is more stable against extreme values, and it can be used in two ways (see below). The index can be written in several ways, but it is frequently written as follows:

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \; i \neq j$$

Null hypothesis H0, states that there is no spatial correlation in the given territory. Vice versa, if $Iw > 0$, then there is a positive autocorrelation, which means that high values are neighbouring high ones and low values are neighbouring low ones. In the case of a negative autocorrelation, the reverse would apply. Depending on the distribution of a spatial variable, the calculation of a median value: neighbour high ones and low values neighbour low ones. Depending on the distribution of a spatial variable, the calculation of a median value is

$$E(I_W) = E(c_W) = -\frac{1}{n-1}$$

and the calculation for the testing statistics is

$$\frac{I_W - E(I_W)}{\sqrt{\text{Var}(I_W)}} \sim \frac{c_W - E(c_W)}{\sqrt{\text{Var}(c_W)}} \sim \text{N}(0, 1)$$

A key element for calculating the indices of spatial autocorrelation is to determine the neighbourhood, i.e. to select spatial entities that are neighbours by *definition*. In the case of this study, the neighbourhood was defined by the edges_corners method, i.e. grids that had a common edge or vertex were always taken

as neighbours. An explanation of this approach can be found in (*Kraus*, 2019).

The Moran's index is a global statistic, which provides no information about the extent of local variation in spatial variability. For that, there are tools that enable us to assess the local level of spatial autocorrelation (LISA) and to measure the intensity and importance of autocorrelation between the value of the variable in a spatial unit and the value of the same variable in neighbouring spatial units. These indicators examine the following two features:

- for each observation they show the intensity of the clustering of similar/opposite values around that observation;
- the sum of local indices at all observations is proportional to the corresponding global index, e.g. to global Moran's I.

In the case of Moran's I, its local value can be written as follows

$$I_i = (y_i - \bar{y}) \sum_j w_{ij} (y_j - \bar{y})$$

and the value of the global index is as follows

$$I_w = \text{konst} \cdot \sum_i I_i$$

where:

- $I_i > 0$ indicates the clustering of similar values (higher or lower than the average for a given neighbourhood), and
- $I_i < 0$ indicates the clustering of different values.

The spatial clustering of similar or different values is observed as follows: as High-High values (HH), Low-Low values (LL), High-Low values (HL), or Low-High (LH) values. If we mean a high value surrounded by other high values or a low value surrounded by other low values then they are referred to as hot spots or cold spots, respectively. If we mean a high value surrounded by low values or a low value surrounded by high values, then these are spatial outliers (*Anselin*, 1995). The significance of each local indicator is based on a spatial distribution of data and statistics that is asymptotically approaching the normal distribution:

$$z(I_i) = \frac{I - E(I_i)}{\sqrt{\text{Var}(I_i)}} \sim \text{N}(0, 1)$$

Since the global rate of spatial autocorrelation (Moran's I) proved to be distinctively higher in the case where the neighbouring municipalities method is used, local rates of Moran's I were further computed only for this method of neighbourhood determination.

RESULTS

A relatively simple model was chosen for the calculation (it is a test), where the output (i.e. perturbed) variable is the number of people who are usually living according to the grid network. This total is information that can be published without restriction. The constraint occurs when it needs to be published in combination with another variable or variables. For the purpose of this test, two variables that enter perturbation were selected: sex and age. As the combination of age, sex, and individual grid units would create too low a frequency, age was transcoded into ten-year groups in line with Eurostat's recommendation: the output is the number of usually living by sex, age, ten-year age groups, and grids. These combinations were then aggregated again into the number usually living according to the grid network and the result was evaluated.

According to the Eurostat methodology, two variants were calculated, which differ by setting the parameters maximum noise D and variance of noise V. D = 3 and V = 1 are settings recommended on the basis of Eurostat testing. Furthermore, in accordance with the Eurostat methodology, version 4 was calculated with values D = 2 and V = 1, i.e. with a lower level of perturbation parameter D but with the same level of noise variance V. Another option is to keep zero values, i.e. grids with a zero number of habitual residents. They are not subject to perturbations.

Cantabular was configured with a perturbation table designed to replicate Eurostat variant 1, but with a reduced cell-key range, compatible with Cantabular. It had a maximum absolute perturbation parameter D of 3 and a noise variance V of approximately 1 (*Cantabular*, 2021).

Recommended statistics were calculated for the difference in the number of usually living between

Table 1 Simple descriptive statistics for absolute difference (AD) - Eurostat solution

Variant 1									
Maximum	Mean	Median	60th pctl	70th pctl	80th pctl	90th pctl	95th pctl	99th pctl	Variance
3	0.41	0	0	1	1	1	2	2	0.39
Variant 4									
2	0.49	0	1	1	1	1	2	2	0.42

Source: Author's calculation.

Table 2 Simple descriptive statistics for absolute difference (AD) - Cantabular solution

Variant 1									
Maximum	Mean	Median	60th pctl	70th pctl	80th pctl	90th pctl	95th pctl	99th pctl	Variance
49	2.82	0	1	3	5	9	13	22	23.05

Source: Author's calculation.

the original and the perturbed value according to the grid network.

The result of the absolute difference (AD) shows that there is a clear difference between the results according to the Eurostat and Cantabular method in the case of maximum, mean and variance. This difference is not so significant for the median and lower percentiles.

The higher maximum value for absolute difference shown in the table above for Cantabular is caused by a query being run at a high level of detail – age by sex by grid square – before the results are then added up at a lower level of detail – total population by grid square – for a comparison with the original unperturbed data.

This has the effect of compounding perturbation because of the loss of additivity in the marginal totals that is inherent in the cell-key method. If the initial query was done at total population by grid square, the maximum absolute difference would be 3, as set in the perturbation configuration. As discussed above, Cantabular does not attempt to restore additivity in order to provide a larger, more flexible range of outputs (*Cantabular*, 2021).

The cumulative distribution function $\boldsymbol{F}_{AD}$ (proportion of cells with an absolute difference less than d was calculated for = 1 to 15. While in the case of results according to the Eurostat methodology there was a complete enumeration in variant 1 for CDF = 3 and in the case of variant 4 even for CDF = 2, the results according to the Cantabular methodology show a gradual and uniform increase in frequencies up to value 15. This follows from a previous finding of a maximum of ***AD***, which was for Eurostat variants 2 and 3, while for Cantabular was 49.

However, in the case of the relative absolute difference ($\boldsymbol{R}_{AD}$), the differences between the Eurostat and Cantabular methodologies are blurred. The maximum $\boldsymbol{R}_{AD}$ reaches the value 3 for both variant 1 of the Eurostat methodology

Table 3 Cumulative distribution function (CDF) for absolute difference – Eurostat solution

Variant 1					Variant 4				
CDF	Frequency	Percent	Cumulative frequency	Cumulative percent	CDF	Frequency	Percent	Cumulative frequency	Cumulative percent
0	42,100	65.67	42,100	65.67	0	38,158	59.53	38,158	59.53
1	18,074	28.19	60,174	93.87	1	20,546	32.05	58,704	91.58
2	3,575	5.58	63,749	99.45	2	5,400	8.42	64,104	100.00
3	355	0.55	64,104	100.00					

Source: Author's calculation.

Table 4 Cumulative distribution function (CDF) for absolute difference – Cantabular solution

Variant 1				
CDF	**Frequency**	**Percent**	**Cumulative frequency**	**Cumulative percent**
0	40,427	50.45	40,427	50.45
1	8,345	10.41	48,772	60.57
2	6,028	7.52	54,800	68.39
3	4,504	5.62	59,304	74.01
4	3,501	4.37	62,805	78.38
5	2,786	3.48	65,591	81.86
6	2,249	2.81	67,840	84.66
7	1,856	2.32	69,696	86.98
8	1,621	2.02	71,317	89.00
9	1,369	1.71	72,686	90.71
10	1,120	1.40	73,806	92.11
11	985	1.23	74,791	93.34
12	856	1.07	75,647	94.41
13	692	0.86	76,339	95.27
14	654	0.82	76,993	96.09
15	3,136	3.91	80,129	100.00

Source: Author's calculation.

Table 5 Simple descriptive statistics for relative absolute difference – Eurostat solution

Variant 1									
Maximum	**Mean**	**Median**	**60th pctl**	**70th pctl**	**80th pctl**	**90th pctl**	**95th pctl**	**99th pctl**	**Variance**
3	0.28	0.06	0.14	0.29	0.50	1.00	1.00	2.00	0.39
Variant 4									
2	0.35	0.11	0.22	0.50	1.00	1.00	1.00	2.00	0.42

Source: Author's calculation.

Table 6 Simple descriptive statistics for relative absolute difference – Cantabular solution

Variant 1									
Maximum	**Mean**	**Median**	**60th pctl**	**70th pctl**	**80th pctl**	**90th pctl**	**95th pctl**	**99th pctl**	**Variance**
3	0.19	0.08	0.12	0.18	0.29	0.50	0.86	1.00	23.05

Source: Author's calculation.

and the Cantabular methodology. Similarly, both methodologies yield completely comparable values for both the mean and the percentile values. This is because the denominator of the indicator contains the numbers of original values, so that even in the case of differences between the original and the perturbed value of higher frequencies, the relative differences decrease.

The CDF results for variable $\boldsymbol{R}_{AD}$ show that a higher degree of agreement between the original and the perturbed value exists at lower CDF_RAD levels for the Eurostat method, but with increasing value the situation rotates and for 0.50 the Cantabular method contains 93 percent of all (cumulative values) and while for 0.50 Eurostat methods 1 and 4 contain, respectively, 88 and 85 percent. The results are therefore similar.

The relative Hellinger distance (HDutility) again shows that both methods yield completely comparable

Table 7 Cumulative distribution function (CDF) for relative absolute difference (RAD) – Eurostat solution

Variant 1					Variant 4				
CDF_RAD	Frequency	Percent	Cumulative frequency	Cumulative percent	CDF_RAD	Frequency	Percent	Cumulative frequency	Cumulative percent
0.02	43,422	67.74	43,422	67.74	0.02	41,238	64.33	41,238	64.33
0.05	2,334	3.64	45,756	71.38	0.05	2,304	3.59	43,542	67.92
0.10	3,028	4.72	48,784	76.10	0.10	3,087	4.82	46,629	72.74
0.20	3,391	5.29	52,175	81.39	0.20	3,563	5.56	50,192	78.30
0.30	2,379	3.71	54,554	85.10	0.30	2,428	3.79	52,620	82.09
0.40	1,508	2.35	56,062	87.45	0.40	1,573	2.45	54,193	84.54
0.50	206	0.32	56,268	87.78	0.50	225	0.35	54,418	84.89
0.99	2,547	3.97	58,815	91.75	0.99	2,568	4.01	56,986	88.90
1.00	5,289	8.25	64,104	100.00	1.00	7,118	11.10	64,104	100.00

Source: Author's calculation.

Table 8 Cumulative distribution function (CDF) for relative absolute difference (RAD) – Cantabular solution

Variant 1				
CDF_RAD	Frequency	Percent	Cumulative frequency	Cumulative percent
0.02	47,077	58.75	47,077	58.75
0.05	6,623	8.27	53,700	67.02
0.10	6,753	8.43	60,453	75.44
0.20	7,021	8.76	67,474	84.21
0.30	4,049	5.05	71,523	89.26
0.40	2,178	2.72	73,701	91.98
0.50	1,029	1.28	74,730	93.26
0.99	3,207	4.00	77,937	97.26
1.00	2,192	2.74	80,129	100.00

Source: Author's calculation.

Table 9 Hellinger distance (HD) and related utility measures – Eurostat solution

HD	HDutility	Max difference	Mean Abs Difference	rootMeanSquare
Version 1				
212.13	0.93	3.00	0.34	0.68
Version 4				
294.33	0.91	2.00	0.41	0.74

Source: Author's calculation.

Table 10 Hellinger distance (HD) and related utility measures – Cantabular solution

HD	HDutility	Max difference	Mean Abs Difference	rootMeanSquare
Version 1				
69.53	0.98	49	2.82	5.57

Source: Author's calculation.

results. For the Eurostat method, this agreement is at the level of 0.93, resp. 0.91 and in the case of the Cantabular method even 0.98. This indicates that there are no statistically significant differences between the original and the perturbed values.

All of the measures of ordinal association indicate a positive association. The resulting association rates are comparable for all three methods and the ASE value indicates that they are statistically significant. Slightly higher values obtained by the Cantabular method suggest in favour of this method of data perturbation.

Gini's concentration coefficient is used in geographic surveys because it overcomes the deficiencies of the coefficient of variation depending on the average and is therefore more appropriate for affecting the variability of asymmetric distributions typical of socio-geographical phenomena (*Netrdová et al.*, 2012). An interesting comparison is the one with the result of the GINI index calculation between the original data and the perturbed data. The results show that the value of the Gini index expresses high inequality among values, but at approximately the same level for the original and the perturbed data.

Previous results showed a statistical evaluation of the results, without questioning whether the original and perturbed values are somehow differently distributed in space. The answer to this question is given by the global and local measures of spatial autocorrelation.

Table 11 Measures of association between original and perturbed data

Statistic	Eurostat – version 1				Eurostat – version 4				Cantabular			
	Value	ASE	95% Confidence limits		Value	ASE	95% Confidence limits		Value	ASE	95% Confidence limits	
Pearson correlation (Rank Scores)	0.730	0.001	0.728	0.731	0.730	0.001	0.729	0.731	0.955	0.000	0.954	0.956
Lambda asymmetric C|R	0.329	0.001	0.327	0.330	0.288	0.001	0.286	0.289	0.155	0.002	0.152	0.159
Lambda asymmetric R|C	0.393	0.001	0.392	0.394	0.268	0.001	0.267	0.269	0.176	0.002	0.173	0.180
Lambda symmetric	0.362	0.001	0.361	0.364	0.277	0.001	0.276	0.279	0.166	0.002	0.163	0.169
Uncertainty coefficient C|R	0.701	0.000	0.701	0.702	0.650	0.000	0.649	0.650	0.628	0.001	0.626	0.630

Source: Author's calculation.
C|R – columns|rows, ASE - asymptotic standard errors (*Stokes et al.*, 2012, p. 125).

Table 12 Gini coefficient for original and perturbed data

Gini coefficient	Original data	Eurostat – version 1	Eurostat – version 4	Cantabular perturbed data
		Perturbed data	Perturbed data	
	0.893	0.895	0.895	0.900

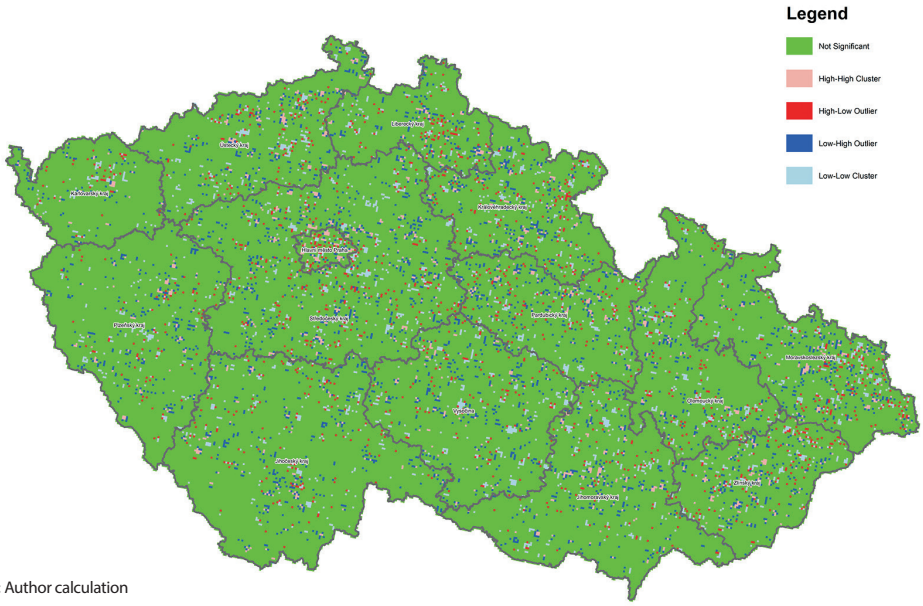
Source: Author's calculation.
C|R – columns|rows, ASE - asymptotic standard errors (*Stokes et al.*, 2012, p. 125).

Table 13 Global Moran's I summary for original and perturbed data

	Original data	Eurostat		Cantabular variant
		Variant 1	Variant 4	
Moran's Index	0.493	0.493	0.492	0.496
Variance	0.000	0.000	0.000	0.000
z-score	277.8	277.8	277.8	280.5
p-value	0.0	0.0	0.0	0.0

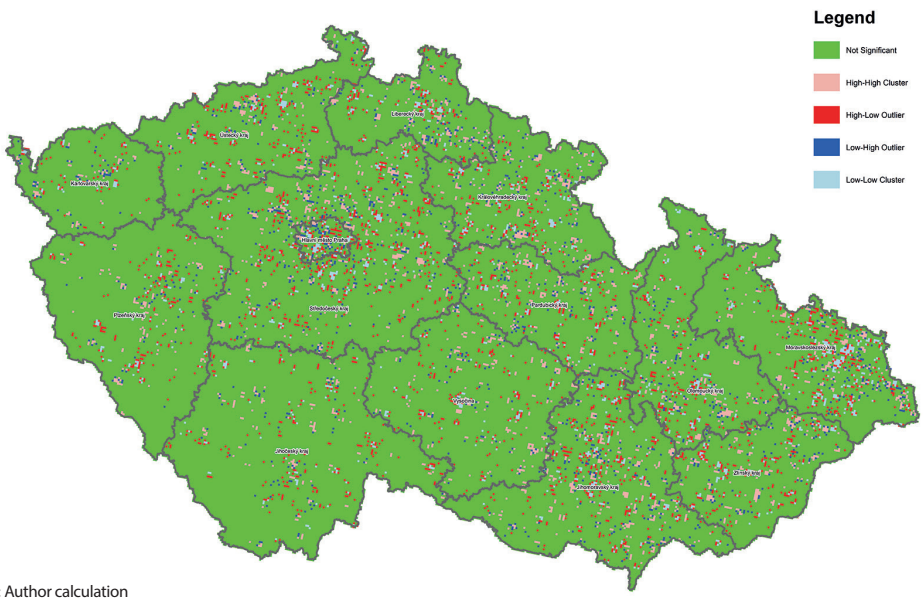
Source: Author's calculation.

Figure 2 Typology of grids according to the difference between the original and the perturbed value calculated according to Eurostat methodology (local Moran's I)



Source: Author calculation

Figure 3 Typology of grids according to the difference between the original and perturbed values calculated according to the Cantabular methodology (local Moran's I)



Source: Author calculation

Note: The above cartograms show that perturbation does not lead to a change in the spatial structure of the observed phenomenon, which in this case is the number of usually living in the individual grids. If there was a change, then the dominant (not significant) value, marked in green, would be replaced in larger areas (i.e. groups of grids) by a different colour than, and thus the structure would be disrupted.
Because this is not the perturbed value of the number of usually living by grids, the derived structures used in this model case (five-year age structures, sex) are guaranteed to yield consistent results (i.e. compared to the original values).

Moran's I calculation was based on the neighbourhood defined by the Edges Corners method, meaning that neighbours are those grids that have either an edge or a corner in common. The results show that the value of the index is practically the same for the original and perturbed data calculated by both the Eurostat and Cantabular methods and differs only to the third decimal place. Given the p-value, the pattern appears to be significantly different from random, and the z-score indicates that all models are very similar.

The local level of spatial autocorrelation (LISA) indicates the local values of Moran's I. This indicator was calculated for the difference between the original and the perturbed value of each grid. From Figures 2 and 3 it is evident that the type Not Significant (the bright green colour) predominates, i.e. perturbation is also a spatially random process that does not change the spatial distribution of usually living.

CONCLUSION

The 2011 Population and Housing Census in the Czech Republic was accompanied by a significant change in the technology used to prepare the census and in the actual course of fieldwork, along with changes in how the data were processed and the outputs were disseminated. Some methodological approaches to processing the data have also changed and are now more aligned with international recommendations. Although a number of changes have been relatively widely discussed in the literature, one type of output remains somewhat overlooked: census results in a grid network.

Working with a network of grids has both advantages and disadvantages, but the main disadvantage is that grids are small territorial units that are often minimally populated. This is mainly a problem in terms of the protection of individual data, which is associated with statistical disclosure control (SDC).

The research question addressed in this paper is whether data protection (perturbation methods) leads to a change in the characteristics of the file either in terms of the statistics of the whole file (i.e. for all grids) or in terms of spatial statistics, which indicate the spatial distribution of the analysed phenomenon.

Two possible solutions to the issue of grid data protection were examined. One comes from the Statistical Office of the European Communities (Eurostat) and the other from Cantabular, which is a product of the Belfast company Sensibile Code Ltd.

In both cases, the data protection solutions are described. One possible solution is to add noise to tabular outputs, using the cell-key method. Tables are produced dynamically from microdata in real-time in response to a user's query and noise is added deterministically based on a computed cell-key and a perturbation table. Zeros can also be perturbed without affecting any structural zeros found in the data for each query.

The starting point for measuring the loss of information due to the use of SDC is the evaluation of frequency tables, i.e. the analysis of the differences between the original and the perturbed value. For perturbative methods, measures of the maximas, means, medians, and some percentiles of absolute differences (AD) and relative differences (RAD) between the original and altered counts in a table and the (squared) differences of the square roots between original and altered counts were calculated.

However, in the case of grids, it was also necessary to focus on spatial measures, which measure the degree of spatial distribution of both the original and the perturbed data sets, e.g. the global autocorrelation rate and the local autocorrelation rates.

The results are based on a relatively simple model for the calculation, where the output (i.e. perturbed) variable is the number of people usually living according to the grid network. The constraint occurs when it should be published in combination with another variable or variables. For the purpose of this test, two variables that enter perturbation were selected: sex and age. As the combination of age, sex, and individual grid units would create too low a frequency, age was transcoded into ten-year groups in line with Eurostat's recommendation: the output is the numbers usually living by sex, age, ten-year age groups, and grids. These combinations were then aggregated again into the number usually living according to the grid network and the result was evaluated.

According to the Eurostat methodology, two variants were calculated, which differ by the parameters set for maximum noise D and the variance of noise V. Another option is to keep zero values, i.e. grids with a zero number of habitual residents. They are not subject to perturbations. Cantabular was configured with a perturbation table designed to replicate Eurostat variant 1, but with a reduced cell-key range compatible with Cantabular.

The results of the descriptive statistics show a difference in the absolute differences compared with the Eurostat methodology, and Cantabular explains the different way of processing microdata. In the case of other statistics, the results are fully comparable.

This paper is devoted to one specific type of census output. The question is to what extent these results are relevant for other types of outputs and in particular for outputs in hypercubes. They differ fundamentally in terms of the number of dimensions (grids have only two dimensions). It would therefore be appropriate to use SDC procedures that allow greater flexibility in defining SDC parameters.

References

- Anselin, L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2), pp. 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.
- Cantabular. 2021. Internal communication with The Sensible Code Co. (Mike Thompson – Aine McGuire), 5. 5. 2021
- Company The Sensible Code. 2020. Press Release. https://sensiblecode.io/resources/case-study-ons.pdf.
- Český statistický úřad (CZSO), n.d. *Pramenné dílo SLDB 2011.* 2013th edition. Prague: Czech Statistical Office, 461 pp.
- Domingo-Ferrer, J. – Franconi, L. (dir.). 2006, *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006, Proceedings*, Berlin Heidelberg, Springer-Verlag, Information Systems and Applications, incl. Internet/Web, and HCI. https://doi.org/10.1007/11930242.
- Domingo-Ferrer, J. – Montes, F. (dir.) 2018. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, Springer International Publishing, Information Systems and Applications, incl. Internet/Web, and HCI. https://doi.org/10.1007/978-3-319-99771-1.
- Dixon, P. M. – Weine, J. – Mitchell-Olds, T. – Woodley, R. 1987. Bootstrapping the Gini Coefficient of Inequality. *Ecology*, 68(5), pp. 1548–1551. https://doi.org/10.2307/1939238.
- Doxsey-Whitfield, E. – MacManus, K. – Adamo, S. B. – Pistolesi, L. – Squires, J. – Borkovska, O. – Baptista, S. R. 2015. Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4. *Papers in Applied Geography*, 1(3), pp. 226–234. https://doi.org/10.1080/23754931.2015.1014272.
- Eurostat. 2015. Centre of Excellence on Statistical Disclosure Control, *CROS - European Commission.* https://ec.europa.eu/eurostat/cros/content/centre-excellence-statistical-disclosure-control-0_en.
- Hundepool, A. – Domingo-Ferrer, J. – Franconi, L. – Giessing, S. – Schulte Nordholt, E.– Spicer K. – de Wolf P.-P. 2012. *Statistical Disclosure Control.* John Wiley & Sons, 270 p. https://doi.org/10.1002/9781118348239.
- Klauda, P. 2011. Site-Oriented Statistics and its Geoinformatic Potential. *Statistika - Economy and Statistics Journal*, 2011(2), pp. 107–110.
- Kraus, J. – Moravec, Š. 2014. Prezentace výsledků SLDB 2011 v síti čtverců – projekt GEOSTAT," *Demografie*, 2014(56), pp. 143–146.
- Lloyd, C. T. – Chamberlain, H. – Kerr, D. – Yetman, G. – Pistolesi, L. – Stevens, F. R. – Gaughan, A. E., Nieves, J. J. – Hornby, G. – MacManus, K. – Sinha, P. – Bondarenko, M. – Sorichetta, A. – Tatem, A. J. 2019. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data*, 3(2), pp. 108–139. https://doi.org/10.1080/20964471.2019.1625151.
- Lloyd, C. T. – Sorichetta, A. – Tatem, A. J. 2017. High resolution global gridded data for use in population studies. *Scientific Data*, 4(1), p. 170. https://doi.org/10.1038/sdata.2017.1.
- Loonis, V. – de Bellefon, M.-P. 2018. *INSEE-EFGS-Eurostat, Handbook of Spatial Analysis.* 2018th edition. INSEE – Eurostat, INSEE, 394 p.
- Martin, D. – Lloyd, C. – Shuttleworth, I. 2011. Evaluation of Gridded Population Models Using 2001 Northern Ireland Census Data. *Environment and Planning A: Economy and Space*, 43(8), pp. 1965–1980. https://doi.org/10.1068/a43485.

- Netrdová, P. – Blažek, J. Aktuální tendence lokální diferenciace vybraných socioekonomických subjektů v Česku: směřuje vývoj k větší mozaikovitosti prostorového uspořádání? (in Czech) *Geografie*, 2012, 3, pp. 266–288.
- SAS. Stat SAS(R) 9.2 User's Guide [online]. 2nd Ed. [cit. 27.04.2021] https://v8doc.sas.com/sashtml/stat/chap28/sect20.htm.
- Shlomo, N. – Young, C. 2006, Statistical Disclosure Control Methods Through a Risk-Utility Framework. In: Domingo-Ferrer, J. – Franconi, L. (eds). *Privacy in Statistical Databases.* https://doi.org/10.1007/11930242_7.
- Shlomo, N. 2010b. *Measurement Error and Statistical Disclosure Control*, 118 p. https://doi.org/10.1007/978-3-642-15838-4_11.
- Stokes M. E. – Davis Ch. S. – Koch G. G. 2012. *Categorical Data Analysis Using SAS. Third Edition*, 3rd edition. Cary, NC, SAS Institute, 590 p.
- Templ M. 2017. Statistical disclosure control for microdata. *Cham: Springer International Publishing.* https://doi.org/10.1007/978-3-319-50272-4.
- Thijs B. – Welch M. n.d. Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro — SDC Practice Guide documentation. https://sdcpractice.readthedocs.io/en/latest/.

JAROSLAV KRAUS

graduated from the Prague University of Economics and Business and completed his PhD studies at the Faculty of Science, Charles University. He works at the Czech Statistical Office in the Department of Population Statistics. He specialises in the issue of population censuses and especially in data processing, the statistical protection of census data, and the spatial analysis of demographic phenomena. He has published several papers in this area in recent years.