*STATISTICS IN TRANSITION new series, June 2025*
*Vol. 26, No. 2, pp. XX–XX,* https://doi.org/10.59139/stattrans-2025-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

# Application of Statistical Disclosure Control methods to protect the confidentiality of the 2020 agricultural census microdata[1]

**Andrzej Młodak**[2], **Tomasz Józefowski**[3]

## Abstract

In this paper, we describe an attempt made to develop an efficient disclosure control algorithm for microdata in a statistical portal used for releasing detailed statistical information at various levels of spatial aggregation. The proposed algorithm is based on perturbative methods, such as microaggregation with Gower's distance for categorical variables and the addition of correlated noise for continuous variables, but it also offers several alternative options in this regard. Moreover, the algorithm can be used to assess the loss of information by measuring distribution disturbances (based on a complex distance that accounts for all measurement scales) and the impact of the Statistical Disclosure Control (SDC) on the strength of correlations between variables (for continuous variables). Through the application of the tools offered by the `sdcMicro` R package, the algorithm was tested using microdata about agricultural farms and farm animals collected in the 2020 Polish Agricultural Census. We present the results of the tests and discuss the main problems and challenges connected with the use of such tools.

**Key words:** Statistical Disclosure Control, perturbative methods, disclosure risk, information loss, agricultural census.

## 1. Introduction

Censuses are the biggest and most informative statistical data collection undertakings. They provide key data about the population, households and farms. This is why they are of particular interest to all groups of users, including government agencies and units of local government administration, policy makers, and various organizations. In other words, the demand for detailed and comprehensive census data is especially high.

Before census data can be safely released, they have to undergo a meticulous process of statistical disclosure control (SDC) to ensure that sensitive information remains confidential. The primary task of every national statistical institute consists in striking an optimal balance between minimizing the risk of disclosure and maximizing the utility of disclosed data

[1]The paper was presented at the International Conference "Privacy in Statistical Databases 2024" (PSD 2024) in Antibes Juan-les-Pins, France, September 25–27, 2024.

[2]Statistical Office in Poznań, Centre for Small Area Estimation; address: Statistical Office in Poznań, Branch in Kalisz, ul. Piwonicka 7-9, 62-800 Kalisz, Poland; e-mail: a.mlodak@stat.gov.pl & University of Kalisz, Interfaculty Department of Mathematics and Statistics, ul. Nowy Świat 4a, 62-800 Kalisz, Poland. ORCID: https://orcid.org/0000-0002-6853-9163.

[3]Poznań University of Economics and Business, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland; e-mail: tomasz.jozefowski@ue.poznan.pl & Statistical Office in Poznań, Centre for Small Area Estimation, ul. Wojska Polskiego 27/29, 60-624 Poznań, Poland. ORCID: https://orcid.org/0000-0001-9485-1946.

(i.e. minimizing the loss of information resulting from SDC measures). Which SDC methods and to what extent they should be used depends, of course, on the format of data publication and their specific characteristics.

In this paper we present and discuss solutions used to test the process of releasing microdata from the Agricultural Census of 2020. This work was part of a broader research project concerning applications of SDC to protect census data (including microdata from the 2021 National Population and Housing Census). The project focused on three different forms of releasing statistical information:

- microdata
- data at 1 km $\times$ 1 km resolution
- hypercubes.

Different methods and tools are required in order to protect data confidentiality in each of these cases. Sets of microdata are collections of information about individual units (e.g. persons); hypercubes are multidimensional tables, while gridded data can be treated as a kind of table (a special type of hypercube), but given very small counts in the majority of the cells, it is possible to apply some methods dedicated to microdata.

The use of SDC methods for census data has been broadly investigated in the literature (cf. e.g. Zayatz (2002), Shlomo et al. (2010), Jansson (2012), Calian (2020), Kraus (2021), SNI et al. (2022) or Muralidhar and Domingo-Ferrer (2023)). The authors of these papers propose various approaches for specific cases with a specific set of parameters. A number of methods have been developed to protect census data, such as Targeted Record Swapping or the Cell Key Method. Moreover, the US Census Bureau adopted a differential privacy approach based on the assumption that a change to one entry in a database only creates a small change in the probability distribution of the outputs (cf. e.g. Abowd (2018)). Since then, this approach has been modified in various ways. For instance, Tran et al. (2024) propose using quantile regression to improve the utility of data protected by differential privacy. Jackson et al. (2024) demonstrate how to apply differential privacy to efficiently protect tabular data using the Poisson synthesis mechanism. It is also becoming increasingly common to rely on neural networks – such as Generative Adversarial Networks, GAN – to generate synthetic data (cf. Yoon et al. (2020)). Innovative approaches have also been proposed to measure the risk of disclosure and information loss. For example, Shlomo (2022) develops distance metrics to compare the overall distributions in the original data versus synthetic data for a particular variable and, more specifically, within equivalence classes based on the Kullback–Leibler distance, the Total Variation and Hellinger's Distance. Shlomo and Skinner (2022) use microdata from a sample survey to infer population parameters when the population is unknown, and estimate the risk of re-identification based on the notion of population uniqueness using probabilistic modelling. A synthetic review of modern concepts in this field can also be found in Templ (2017). Młodak (2020) proposes a new method of assessing information loss in terms of distribution disturbance based on the idea of the Gower distance, where the cyclometric function is used in the partial distance for continuous variables, and a method of computing information loss on regarding relationships between continuous variables using an inverse correlation matrix. An improved

version of this method is described in Młodak et al. (2022), which also presents a method of assessing external (ex post) disclosure risk – when it is assumed that the end user has access to an alternative data set containing information that can be linked with statistical data in order to identify a unit.

We show the particular nature of the data from Polish censuses and propose optimum SDC methods to protect microdata from the 2020 Agricultural Census. The purpose of the work conducted during this study was to prepare an algorithm for protecting microdata to be released in the Geostatistics Portal maintained by Statistics Poland. The microdata were used primarily to test the algorithm's efficiency, but, ultimately, they will be also uploaded to the portal. The work was conducted between 1st July 2018 to 31st October 2022 as part of the project "Spatial statistical data in the state information system", implemented under the Operational Programme Digital Poland within Priority axis II – "E-administration and open government" Measure 2.1. "High availability and quality of public e-services". All actions were financed from the European Regional Development Fund (ERDF). The main contractor was a consortium of companies with a documented track record of data processing and statistical analysis. The underlying assumptions of the SDC process were specified by Statistics Poland.

The final algorithm was based on the perturbative methods, such as microaggregation and noise addition. The efficiency of the methods we applied was assessed with measures of disclosure risk (based on $k$-anonymity and the concept of individual and global risk) and information loss (above all, these proposed by Młodak et al. (2022)). The methods we chose and their parameters as well as the most important problems to be solved in the future are presented and discussed below.

The paper is organised as follows. Section 2 presents basic assumptions of the Geostatistics Portal and the project "Spatial statistical data in the state information system", especially with respect to microdata. Section 3 describes the set of microdata from the 2020 Agricultural Census and their basic characteristics. Section 4 contains a description of various methods of statistical disclosure control and dedicated software. Section 5 contains the most important results and addresses problems encountered during the study. Section 6 includes the key conclusions.

## 2. The release of microdata via the Geostatistics Portal

The Geostatistics Portal of Statistics Poland (https://portal.geo.stat.gov.pl/en/home/)[4] was created to satisfy the demand for detailed and high-quality data at various levels of spatial aggregation, enabling users to conduct their own studies and analyses and present results in their preferred form (tabular or graphical). The Portal was developed as part of the Spatial Statistical Data project, with a goal of expanding the scope and availability of statistical information and geostatistical analysis methods that rely on publicly available statistical data.

Before any innovations were implemented, user needs and current limitations of the Geostatistics Portal were analyzed. One of the expectations was that the portal should

[4] Some information given in the following paragraphs is based on the description included on the webpage.

provide a wide range of possible tools for analyzing the spatial distribution of various socio-economic phenomena in specific areas and with a high level of detail and precision. Therefore, an option of analyzing gridded microdata (1 km $\times$ 1 km grid) was added and tools were provided to enable users to independently conduct advanced statistical analyses, especially at various levels of spatial aggregation.

The improved functionalities of the Portal include tools for statistical analyses at any level of spatial aggregation, possibility of combining statistical data with users' own data, geocoding user objects used for geostatistical analyses, using exploratory analyses of spatial data based on statistical information, performing geostatistical modelling and the option of enriching users' own content with geostatistical information and analyses.

In summary, the main outcome of the project are a number of publicly available e-services:

- the ability to access statistical information collected in the Portal from a remote computer and perform advanced spatial analyses using available data and metadata (users can select data, area, visualization method and method parameters). The user can generate an analysis of a given spatial area (also at 1km x 1km resolution) and present its results in a choropleth map or various types of editable cartodiagrams, which can show the variability of statistical data over time,

- the ability to access the portal from a mobile device (using an Android and iOS application). The aforementioned functionalities are adapted to being displayed on smaller screens,

- the ability to perform exploratory analyses of spatial data using statistical information stored in the Portal; using the available tools users can examine the spatial distribution of selected variables and determine spatial connections, interdependencies and identify clusters. A variety of descriptive statistics and statistical methods are available (e.g. central tendency statistics, dispersion statistics, measures of asymmetry and concentration, variable correlation analyses, etc.). It is also possible to perform cluster analysis and check spatial autocorrelation and similarity of objects,

- the ability to conduct analyses involving geostatistical modelling. e.g. to generalize/ estimate results based on a random sample to other surveyed units or the population of these units. Users can create models and apply a probabilistic model for statistical inference (estimation) concerning values of the response variable based on results of a random sample survey and the assumed probability distribution. There are statistics and tests that can be used to verify the quality of these models as well as some spatial interpolation and imputation methods,

- the ability to enhance user's own content with geostatistical information and analyses provided by the Portal (semantic access to documents related to the analytical work and the ability to supplement user's own text-based content with graphical elements). Additionally, advanced users can use programming languages to access the Portal's databases via the API.

All of these functionalities can support users in decision-making processes related to statistical and spatial information and enable them to benefit from spatial and data mining analyses, either in the context of business activity, or in policy-making by government and local government administration, or in scientific research.

However, before any such detailed statistical data can be released to enable advanced analyses, they have to undergo statistical disclosure control to protect data confidentiality. Apart from satisfying legal requirements, it is necessary to apply additional tools to minimize the risk of potential identification of individual units and unauthorized disclosure of sensitive information about them. Given the complexity of information provided to Portal users, the kind data from other sources they may have access to and the sophistication of their analyses, the SDC process should be conducted thoroughly by competent staff.

Outputs of any analyses conducted in the Portal using data designated as protected (i.e. from internal statistical databases), including map visualizations, have to be checked in terms of primary confidentiality. According to this requirement, values of aggregates can only be displayed (visualized) if they contain a sufficiently large number of units (data records) – at least 3 (it is the fundamental rule established in the Polish Act on Official Statistics), and, in some cases, at least 10. However, in some situations aggregate values suppressed to protect statistical confidentiality could be recalculated by the user on the basis of correlations between results of various analyses (queries). For example, if a higher-order aggregate consists of several lower-order groupings and the value of only one of them is hidden (because it would violate the statistical confidentiality), the hidden value can be determined by subtracting the sum of the displayed components (lower-order groupings) from the value of the higher-order aggregate. Such situations require secondary confidentiality. Normally, this is achieved by additionally suppressing aggregates which apparently (from the point of view of primary confidentiality) do not violate the protection rules, but allow the protected values to be recalculated through the use of indirect dependencies. Because the tools available in the Portal system are flexible and diverse, they enable users to analyze and aggregate data in any way and not be limited to pre-defined formats, it is not possible to create algorithms that will reliably control secondary confidentiality at the stage of presenting analytical results/ data summaries by hiding appropriate aggregates.

For this reason, any analyses based on a protected set of data that are to be released to external users are not performed on the original set of microdata, but on a set of data subjected to data distortion techniques designed to protect statistical confidentiality. Therefore, although users do not get direct access to unit-level data in the system (they can only see aggregates containing at least the minimum number of units), because of the flexibility offered by the aggregation tools which are associated with a a high risk of recalculating information about individual units based on the dependencies between the data, they can work (i.e. perform self-defined aggregations) only on datasets that have been disturbed by appropriate SDC methods. In this way, even if the user is able to recalculate values pertaining to individual units (records) in the disturbed set, this should not result in the disclosure of actually protected information, unless the user relies on individual data from other studies (e.g. registers of labor offices and the Labour Force Survey). Then, the risk of revealing sensitive information by linking relevant records from different sources may increase. In such situations, it may be necessary to carry out an additional - joint -

verification of the provided files in terms of statistical confidentiality. This will also be necessary, if, in the future, users of the system are able to use their own, external data sources.

In summary, under the adopted approach, only disturbed sets of publicly available unit-level data can made available for analysis. To implement this form of protection, the project team created a parameterized script to perform perturbations involving SDC methods.

Perturbation cannot be performed automatically. Each set of statistical data to be released to external users through the system must be perturbed separately. The disruption process (which may have to be repeated in the event of data update) must be performed by an analyst with a knowledge of the specific dataset and SDC methods. In each case the operation involves creating an appropriate script based on the template provided by the contractor, who should define the role of individual variables of the input set in the disturbance process and the method parameters.

Therefore, the SDC process in the system is enabled by an R script, which relies on functions implemented in the `sdcMicro` package (Templ et al. (2015)). The functions are used to apply specific perturbative methods and control the disclosure risk and information loss. Although the script relies on two main families of perturbations, it can be adapted to include other methods, if necessary. The following sections describe the data used for testing and details of the script.

## 3. Microdata from the 2020 Agricultural Census

Our analysis was based on a dataset containing microdata collected during the Agricultural Census conducted in Poland between 1st September to 30th November 2020, with reference to 1st June 2020. The data are to be made available through the Geostatistics Portal, and, in other forms, to all interested persons, especially for scientific purposes. So they will have to be perturbed to prevent potential unit identification and disclosure of its sensitive information. The set in question contained 1,317,400 records and 81 variables. The following 12 variables describe the main features of farms:

- NR_GOS – farm ID,
- SP – legal status,
- Wo_SG – the province where the farm is located;
- Pow_SG – the district (LAU 1 unit) where the farm is located;
- Gm_SG – the commune (LAU 2 unit) where the farm is located;
- KTS1_SG – the macroregion (NUTS 1) where the farm is located;
- KTS3_SG – the region (NUTS 2) where the farm is located[5],
- KTS4_SG – subregion (NUTS 3) where the farm is located,

[5] In Poland regions coincide with the provinces except for the Mazowieckie Province, which is divided into two regions: the City of Warsaw and the rest of the province.

- UG2w – total land area,
- UG2a – area of agricultural land,
- UG_W1 – area of arable land,
- UG_W2 – area of permanent grassland.

The remaining 69 variables describe various aspects of the livestock population. They are presented in Table 1.

**Table 1.** Variables describing the livestock population in the analyzed dataset

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| ZW1 | Breeding of farm animals (yes/no) | ZW45b | Number of laying hens for the production of table eggs |
| ZW2 | Cattle breeding (yes/no) | ZW45c | Number of laying hens for the production of hatching eggs |
| ZW3w | Total cattle population | ZW45d | Number of turkeys |
| ZW3a | Number of bulls under 1 year of age | ZW45e | Number of geese |
| ZW3b | Number of heifers under 1 year of age | ZW45f | Number of ducks |
| ZW3c | Number of bulls aged 1 to 2 years (except for exactly 2-year-old bulls) | ZW45g | Number of remaining poultry |
| ZW3d | Number of heifers aged 1 to 2 years (except for exactly 2-year-old heifers) | ZW45h | Number of ostriches |
| ZW3e | Number of male cattle aged 2 years and over | ZW47 | Number of horses |
| ZW3f | Number of heifers aged 2 years and over | ZW47a | Number of horses three years old and over |
| ZW3g | Number of dairy cows | ZW48 | Total number of rabbits kept for meat |
| ZW3h | Number of other cows | ZW48a | Number of female rabbits capable of breeding |

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| ZW34 | Farm breeding pigs (yes/no) | ZW49 | Number of other fur animals (including fur rabbits) |
| ZW35w | Total pig population | ZW49a | Number of remaining female fur animals |
| ZW35a | Number of piglets weighing up to 20 kg | ZW50 | Number of bee trunks |
| ZW35b | Number of weaners weighing 20-50 kg | ZW51 | Number of remaining animals |
| ZW35c | Number of breeding boars | ZW51a | Number of deer animals |
| ZW35d | Number of pregnant sows | ZW_W1_3 | Number of calves under 1 year of age |
| ZW35e | Number of sows pregnant for the first time | ZW_W2_3 | Number of cattle aged 1-2 years |
| ZW35f | Number of remaining sows (loose, not pregnant) | ZW_W3_3 | Number of cattle aged 2 years and over |
| ZW35g | Number of gilts has never been bred | ZW_W4_3 | Total number of cows |
| ZW35h | Number of pigs for fattening | ZW_W5_3 | Number of female cattle aged 2 years and over |
| ZW40 | Sheep breeding (yes/no) | ZW_W1_35 | Number of pigs for breeding weighing 50 kg and more |
| ZW41w | Total number of sheep | ZW_W2_35 | Total number of breeding sows |
| ZW41a | Number of sheep lambs | ZW_W1_41 | Total number of sheep ewes |
| ZW41b | Number of sheep ewes used for milk production | ZW_W2_41 | Total number of adult sheep |
| ZW41c | Number of sheep ewes used in other directions | ZW_W1_45 | Total number of chicken poultry |
| ZW41d | Number of remaining adult sheep | ZW_W2_45 | Total number of laying hens |
| ZW42 | Goat breeding (yes/no) | ZW_W_SD | Animal population in LSUs[a] |
| ZW43w | Total goat population | ZW_W1_SD | Number of cattle in LSUs |
| ZW43a | Number of female goats one year old and older | ZW_W2_SD | Number of pigs in LSUs |

| Symbol | Description | Symbol | Description |
|---|---|---|---|
| ZW43b | Number of female goats used for milk production | ZW_W3_SD | Number of sheep in LSUs |
| ZW43c | Number of remaining goats | ZW_W4_SD | Number of goats in LSUs |
| ZW44 | Poultry breeding (yes/no) | ZW_W5_SD | Number of poultry in LSUs |
| ZW45w | Total poultry population | ZW_W6_SD | Number of rabbits in LSUs |
| ZW45a | Number of broiler chickens | | |

[a] The livestock unit, abbreviated as LU (or sometimes as LSU - Livestock Standard Unit), means a standard measurement unit that allows the aggregation of various categories of livestock (various species, sex and age) in order to enable them to be compared. Data on animals are converted into livestock units using the following coefficients: equidae – 0.80, young cattle aged less than 1 year old (calves) – 0.40, male bovines aged between 1 and 2 years – 0.70, female bovines aged between 1 and 2 years – 0.70, male bovines aged 2 years and over – 1.00, heifers of bovines aged 2 years and over – 0.80, dairy cows – 1.00, other cows (sucklers) – 0.80, sheep – 0.10, goats – 0.10, piglets with a live weight of less than 20 kg – 0.027, breeding sows with a live weight of 50 kg or more – 0.50, other pigs (young pigs with a live weight of 20 kg or more but less than 50 kg, breeding boars and fattening pigs with a live weight of 50 kg and more) – 0.30, broilers of chickens – 0.007, laying hens – 0.014, other poultry (ducks, turkeys, geese, domestic quails, guinea-fowls, and other poultry but apart from ostriches) – 0.030, ostriches – 0.35 and female of rabbits – 0.020. The reference unit used for the calculation of livestock units (=1 LSU) is the grazing equivalent of one adult dairy cow producing 3 000 kg of milk annually, without additional concentrated foodstuffs.

Source: Based on the metadata for the dataset and information provided by Statistics Poland (https://stat.gov.pl/en/metainformation/glossary/terms-used-in-official-statistics/1394,term.html).

14 of the above variables are categorical (NR_GOS, SP, Wo_SG, Pow_SG, Gm_SG, KTS1_SG, KTS3_SG, KTS4_SG, ZW1, ZW2, ZW34, ZW40, ZW42 and ZW44). The remaining 67 variables are numerical.

The following 26 variables are derived from 55 primary ones: (KTS1_SG, KTS3_SG, KTS4_SG, ZW3w, ZW35w, ZW41w, ZW43w, ZW45w, ZW_W1_3, ZW_W2_3, ZW_W3_3, ZW_W4_3, ZW_W5_3, ZW_W1_35, ZW_W2_35, ZW_W1_41, ZW_W2_41, ZW_W1_45, ZW_W2_45, ZW_W_SD, ZW_W1_SD, ZW_W2_SD, ZW_W3_SD, ZW_W4_SD, ZW_W5_SD, ZW_W6_SD).

These facts were taken into account when planning the SDC process. The next section contains a description of how this information was used to determine the set of quasi-identifiers to be protected and choose appropriate SDC methods.

## 4. Methods and tools of statistical disclosure control

The main problem in defining successive steps of the SDC process was to identify a set of quasi-identifiers that need to be protected. First, the 26 derived variables were excluded from further analysis because any perturbations applied to these variables could cause significant deviations from their original dependencies on primary variables. For instance, the value of ZW43w is the sum of the values of ZW43a, ZW43b and ZW43c. Hence, additivity of these cells should be retained in the safe dataset. Therefore, values of

these derived variables should be re-calculated *ex post*, i.e. after the whole SDC process has been completed. Of course, deviations on particular values of the variables on the basis of which a given derived variable is obtained can accumulate in this way. However, variables are derived at the level of units (not aggregated data, as in tables), so the final accumulation of deviations should be rather low.

There is a group of key quasi-identifiers that describe a given unit's geographical location. These are: Wo_SG, Pow_SG and Gm_SG. Since each of these variables contains only unit codes for a given level (the codes do not contain symbols denoting higher level units) the exact location can only be obtained only by concatenating codes in Wo_SG, Pow_SG and Gm_SG. However, since we allow record swapping between communes (LAU2), we have replaced Wo_SG and Pow_SG by their concatenation, denoted as GEO_ID.

Thus, the variables under analysis are: NR_GOS, GEO_ID, Gm_SG, UG2w, UG2a, ZW1, ZW2, ZW3a, ZW3b, ZW3c, ZW3d, ZW3e, ZW3f, ZW3g, ZW3h, ZW34, ZW35a, ZW35b, ZW35c, ZW35d, ZW35e, ZW35f, ZW35g, ZW35h, ZW40, ZW41a, ZW41b, ZW41c, ZW41d, ZW42, ZW43a, ZW43b, ZW43c, ZW44, ZW45a, ZW45b, ZW45c, ZW45d, ZW45e, ZW45f, ZW45g, ZW45h, ZW47, ZW47a, ZW48, ZW48a, ZW49, ZW49a, ZW50, ZW51 and ZW51a.

Categorical variables were perturbed using microaggregation based on Gower's distance (first described in a PhD thesis by Kowarik (2015) and later also by Templ (2017)). In this method records are combined to form a number of groups. Then, the true value of each sensitive attribute is replaced by a value representing a certain measure of central tendency of this attribute (e.g. mode or mean) for the group a given record belongs to. Groups are formed using a criterion of maximum similarity. Gower's distance is used to compute the distance between any two records, taking into account all measurement scales of variables. Clusters for which microaggregation was to be conducted were established using the variable GEO_ID. Therefore, microaggregation was performed within districts (LAU 2). The Gower's distance was computed using the following variables: UG2w, UG2a, ZW1, ZW2, ZW34, ZW40, ZW42 and ZW44. The mechanism of microaggregation was defined by the `maxCat` function, i.e. the level with the most occurrences is normally chosen or the selection is random if the maximum is not unique. The aggregation level was adjusted for the properties of the analyzed dataset. It is an efficient method of perturbing variables whose values are expressed on various measurement scales, since it offers several possibilities of choosing the form of perturbation and its parameters and its results are easy to interpret. These features give it an advantage over other methods[6]. On the other hand, the method of perturbing continuous variables was chosen because it ensures that relationships between them are retained as much as possible and it reduces the impact of outliers better than many other approaches, which is consistent with basic expectations of users of disclosed data. Of course, the method is slightly sophisticated (but its results are easy to interpret) and might flatten the original distributions (which can be controlled to some extent).

Continuous variables were perturbed using correlated noise addition. The approach involves adding random values selected from a continuous distribution while preserving

[6]In the general script, an alternative use of post-randomization (PRAM) for perturbing categorical variables is available. However, the efficient setting of necessary entries in the transition matrix is more difficult.

the structure of covariances of the original variables and assuring, by way of additional transformations, that the sample covariance matrix of the suppressed variables is an unbiased estimator for the covariance matrix of the original variables (cf. e.g. Kim (1986) or Brand (2002)). The basic parameter $\delta$ and the amount of noise were optimized in a series of trials.

However, the algorithm offers the possibility of using other perturbation methods that are better suited to data with different properties or different user expectations. These other options include post-randomization for categorical variables and microaggregation for continuous variables. The algorithm is an R script and relies on functions from the `sdcMicro` package. To be more precise, the function `microaggrGower` was used to apply microaggregation based on the Gower distance to categorical variables. Noise was added to continuous variables using the `addNoise` function. The risk of disclosure was computed by setting relevant parameters of the `sdcMicroObj` object. The `IL_variables` function was used to assess information loss regarding the distribution and the `IL_correl` function was applied to estimate information loss.

## 5. Results and problems encountered during the exercise

The algorithm took almost 35 hours to complete its run, which mainly resulted from the large number of variables, the complexity of the script and the limitations of the computational environment.

The dataset under analysis contains a few categorical variables. Categories with smallest frequency appear in more than 6 thousands records. Therefore, it is not surprising that the $k$-anonymity rules for $k = 2,3$ and 5 are practically not violated (it is, of course, not a rule; however, the higher the frequency of the "smallest" categories, the lower the probability that the rare combinations occur). Therefore, the risk associated with categorical variables is negligible.

The situation looks very different for the continuous variables. In this case, the risk of disclosure is assessed using the basic function implemented in the sdcMicro package and described by Templ (2017). The function reports the percentage of observations falling within an interval centered on its masked value, whereas the upper bound of such an interval corresponds to the worst case scenario in which an intruder is sure that each nearest neighbor is indeed the true link. The function compares data before and after the SDC process. For raw data the risk – by definition expressed as a percentage – is always in the range between 0% and 100%. The computation showed that after the SDC process the risk interval ranged from [0.00%,100.00%] to [0.00%,0.00%]. Thus, the protection is ideal.

To get a full picture of the efficiency of the SDC process, it is necessary to measure the loss of information. In this experiment, it was assessed in two ways:

- by measuring the distribution disturbance,
- by measuring the disturbance of correlations between the variables.

The measure of distribution disturbance was proposed by Młodak (2020), improved by Młodak et al. (2022) and implemented in the `sdcMicro` package as the `IL_variables` function. It is based on Gower's distance between original and perturbed values and is

defined as the sum of partial distances. In the case of nominal variables, these partial distances amount to 0 if they are the same and 1 otherwise; in the case of ordinal variables, they are equal to the normalized number of categories by which the compared values differ, and in the case of continuous variables, they are computed using the cyclometric function. This measure takes values from [0,1]. The larger the value of the measure, the bigger the loss of information. Information loss can be measured both at the global level and for particular variables. Table 2 shows information loss computed for particular variables.

**Table 2.** Information loss for particular variables (in %)

| Variable | Information loss | Variable | Information loss |
|---|---|---|---|
| NR_GOS | 0.0 | ZW41c | 56.3 |
| Gm_SG | 0.0 | ZW41d | 39.0 |
| UG2w | 92.5 | ZW42 | 0.1 |
| UG2a | 92.1 | ZW43a | 32.2 |
| ZW1 | 0.0 | ZW43b | 24.7 |
| ZW2 | 0.0 | ZW43c | 8.1 |
| ZW3a | 64.7 | ZW44 | 0.0 |
| ZW3b | 66.0 | ZW45a | 99.8 |
| ZW3c | 59.6 | ZW45b | 99.9 |
| ZW3d | 62.1 | ZW45c | 99.2 |
| ZW3e | 27.3 | ZW45d | 99.4 |
| ZW3f | 49.3 | ZW45e | 98.1 |
| ZW3g | 77.2 | ZW45f | 98.6 |
| ZW3h | 47.8 | ZW45g | 94.9 |
| ZW34 | 0.0 | ZW45h | 3.4 |
| ZW35a | 95.9 | ZW47 | 32.8 |
| ZW35b | 96.1 | ZW47a | 21.6 |
| ZW35c | 9.1 | ZW48 | 89.2 |
| ZW35d | 88.5 | ZW48a | 64.2 |
| ZW35e | 68.6 | ZW49 | 99.5 |
| ZW35f | 82.0 | ZW49a | 97.4 |

| Variable | Information loss | Variable | Information loss |
|---|---|---|---|
| ZW35g | 74.4 | ZW50 | 75.7 |
| ZW35h | 96.5 | ZW51 | 82.3 |
| ZW40 | 0.1 | ZW51a | 39.6 |
| ZW41a | 49.0 | GEO_ID | 0.0 |
| ZW41b | 42.7 | GEO_ID_G | 0.0 |

Source: Results obtained by applying the `IL_variables()` function from the `sdcMicro` package.

The variable GEO_ID_G was created for technical reasons by concatenating symbols for province, district and commune. The overall information loss amounts to 53.8%. As can be seen, the level of information loss for particular variables varies greatly. Of course, some variables (e.g. ID_GOS or GEO_ID) could not be changed because of the underlying assumptions of the SDC process. Nevertheless, information loss for the remaining ones varies significantly – from 0.0 to as much as 99.9%. This may be the result of adjustments in the amount of correlated noise in the case of the continuous variables and the fact that some perturbed values may go beyond the range defined for a given variable (and hence some *ex post* corrections will be necessary). On the other hand, however, the distance component for continuous variables (based on the cyclometric function – arcus tangent) tends to take values close to 1 (100%) for larger differences between original and perturbed values. As a result, information loss can be overestimated. On the other hand, information loss can also be overestimated when the original range of values is exceeded as a result of perturbations. As we have noted in Section 6, these inconveniences can be corrected *ex post*, which should reduce this problem. But such overestimation could be helpful when identifying problem areas in the SDC process.

Information loss has some impact on the descriptive statistics of the analyzed variables. Table 3 shows the mean, median and third quartile of primary continuous variables before and after the SDC process.

**Table 3.** Basic descriptive statistics for primary continuous variables before and after the SDC process

| Variable | Original | | | After SDC | | |
|---|---|---|---|---|---|---|
| | mean | median | 3rd quartile | mean | median | 3rd quartile |
| UG2w | 12.6530 | 5.6300 | 11.8600 | 20.5326 | 10.4000 | 32.1400 |
| UG2a | 11.3503 | 4.6900 | 10.2800 | 18.8962 | 9.2500 | 29.5400 |
| ZW3a | 0.6535 | 0.0000 | 0.0000 | 1.8686 | 0.0000 | 3.0000 |
| ZW3b | 0.6582 | 0.0000 | 0.0000 | 1.9537 | 0.0000 | 3.0000 |

| Variable | Original | | | After SDC | | |
|---|---|---|---|---|---|---|
| | mean | median | 3rd quartile | mean | median | 3rd quartile |
| ZW3c | 0.6833 | 0.0000 | 0.0000 | 1.6146 | 0.0000 | 2.0000 |
| ZW3d | 0.6492 | 0.0000 | 0.0000 | 1.7089 | 0.0000 | 3.0000 |
| ZW3e | 0.1009 | 0.0000 | 0.0000 | 0.3713 | 0.0000 | 1.0000 |
| ZW3f | 0.1628 | 0.0000 | 0.0000 | 0.8382 | 0.0000 | 1.0000 |
| ZW3g | 1.6839 | 0.0000 | 0.0000 | 4.0118 | 1.0000 | 6.0000 |
| ZW3h | 0.1972 | 0.0000 | 0.0000 | 0.8302 | 0.0000 | 1.0000 |
| ZW35a | 1.7826 | 0.0000 | 0.0000 | 27.8934 | 1.0000 | 46.0000 |
| ZW35b | 2.5555 | 0.0000 | 0.0000 | 30.1932 | 1.0000 | 49.0000 |
| ZW35c | 0.0107 | 0.0000 | 0.0000 | 0.0993 | 0.0000 | 0.0000 |
| ZW35d | 0.4303 | 0.0000 | 0.0000 | 7.4268 | 0.0000 | 12.0000 |
| ZW35e | 0.0784 | 0.0000 | 0.0000 | 1.7361 | 0.0000 | 3.0000 |
| ZW35f | 0.1875 | 0.0000 | 0.0000 | 3.9925 | 0.0000 | 7.0000 |
| ZW35g | 0.0396 | 0.0000 | 0.0000 | 2.3313 | 0.0000 | 4.0000 |
| ZW35h | 3.4979 | 0.0000 | 0.0000 | 34.7676 | 2.0000 | 56.0000 |
| ZW41a | 0.0531 | 0.0000 | 0.0000 | 0.7486 | 0.0000 | 1.0000 |
| ZW41b | 0.0394 | 0.0000 | 0.0000 | 0.5812 | 0.0000 | 1.0000 |
| ZW41c | 0.0910 | 0.0000 | 0.0000 | 1.0270 | 0.0000 | 2.0000 |
| ZW41d | 0.0433 | 0.0000 | 0.0000 | 0.5094 | 0.0000 | 1.0000 |
| ZW43a | 0.0291 | 0.0000 | 0.0000 | 0.3814 | 0.0000 | 1.0000 |
| ZW43b | 0.0158 | 0.0000 | 0.0000 | 0.2702 | 0.0000 | 0.0000 |
| ZW43c | 0.0119 | 0.0000 | 0.0000 | 0.0911 | 0.0000 | 0.0000 |
| ZW45a | 106.7432 | 0.0000 | 0.0000 | 1258.1000 | 28.0000 | 2003.0000 |
| ZW45b | 34.6117 | 0.0000 | 7.0000 | 2200.8400 | 19.0000 | 3693.0000 |
| ZW45c | 7.3861 | 0.0000 | 0.0000 | 203.2564 | 1.0000 | 334.0000 |
| ZW45d | 13.3664 | 0.0000 | 0.0000 | 248.3547 | 4.0000 | 403.0000 |
| ZW45e | 4.2672 | 0.0000 | 0.0000 | 69.5504 | 1.0000 | 113.0000 |
| ZW45f | 4.4782 | 0.0000 | 0.0000 | 96.2474 | 1.0000 | 158.0000 |

| Variable | Original | | | After SDC | | |
|---|---|---|---|---|---|---|
| | mean | median | 3rd quartile | mean | median | 3rd quartile |
| ZW45g | 0.6233 | 0.0000 | 0.0000 | 20.7730 | 0.0000 | 35.0000 |
| ZW45h | 0.0016 | 0.0000 | 0.0000 | 0.0358 | 0.0000 | 0.0000 |
| ZW47 | 0.1188 | 0.0000 | 0.0000 | 0.4614 | 0.0000 | 1.0000 |
| ZW47a | 0.0694 | 0.0000 | 0.0000 | 0.2764 | 0.0000 | 0.0000 |
| ZW48 | 0.5543 | 0.0000 | 0.0000 | 8.0797 | 0.0000 | 14.0000 |
| ZW48a | 0.1079 | 0.0000 | 0.0000 | 1.4254 | 0.0000 | 2.0000 |
| ZW49 | 3.3246 | 0.0000 | 0.0000 | 297.2698 | 1.0000 | 498.0000 |
| ZW49a | 0.6726 | 0.0000 | 0.0000 | 46.6540 | 0.0000 | 78.0000 |
| ZW50 | 0.4925 | 0.0000 | 0.0000 | 2.8839 | 0.0000 | 4.0000 |
| ZW51 | 0.0873 | 0.0000 | 0.0000 | 3.9979 | 0.0000 | 7.0000 |
| ZW51a | 0.0184 | 0.0000 | 0.0000 | 0.5001 | 0.0000 | 1.0000 |

Source: Results obtained using the SAS Studio software.

As one can see, in most cases the SDC process did not significantly change the presented statistics. Moreover, the original first quartile was 0 except for UG2w (2.8600) and UG2a (2.320), whereas after perturbation this quartile for all variables amounted to 0. However, for some variables – e.g. ZW35a, ZW35b, ZW45a, ZW45b and ZW45c – the differences are more significant. This situation can be due to a large degree of variation in relevant data across various farms and spatial areas, which can have an impact on the noise distribution adjusted to preserve the correlation, according to our assumptions.

The loss of information resulting from the disturbance of correlations between variables, which reflects the degree to which relationships between variables have been preserved, is measured using the approach developed by Młodak (2020), improved by Młodak et al.(2022) and implemented in the `sdcMicro` package as the `IL_correl` function. It is based on distances of normalized sums of diagonal entries of an inverse correlation matrix and takes values from [0,1] (again, the larger the value, the bigger the loss). In the analyzed situation the measure amounts to 7.9%. Therefore, the loss of information about relationships is small. This is largely the result of using correlated noise. Thus, in this respect, SDC seems to be fully efficient.

## 6. Final conclusions

Statistical disclosure control is necessary to ensure the safe and efficient disclosure of statistical information. It is worth emphasizing that without the use of these methods, much

of statistical data would either have to remain unavailable to end users or would largely be useless.

The above exercise indicates that perturbative SDC methods can be very useful, especially if most variables in a given dataset are numerical. As a result, the risk of disclosure associated with these variables is significantly reduced. If there are few categorical variables, then the risk of disclosure associated with them tends to be low (or even negligible).

However, the risk of disclosure is reduced at the cost of some information loss. Perturbations introduced in some variables result in large differences between their original distributions and those after the SDC process. This happens because perturbations cannot preserve some features of the original variables resulting from their definitions, e.g. the range of permitted values. Therefore, additional corrections may be required. On the other hand, in the finally disclosed dataset secondary variables have to be determined (to avoid violations of additivity or related rules, it is reasonable to omit them in the SDC process and to compute them again after it is over), which can have some (rather moderate in the case of microdata) impact on the final information loss.

The use of correlated noise in relation to continuous variables with appropriately chosen parameters results in a small loss of information about relationships between them. So, it is a very important aspect of disclosed data. When appropriate summations of continuous variables are performed to derive secondary variables, these interactions should not be violated.

The algorithm can be a good tool for performing SDC on microdata. However, its application reveals the whole complexity of the process, especially as regards steps that have to be taken before and after the perturbation procedure in order to obtain an output that is efficiently protected and simultaneously sufficiently useful for its users. Thus, each stage of this procedure should be treated with equal care. Of course, it is possible to consider some dynamic methods of protecting data confidentiality. Data in the geostatistics portal will be available as microdata, so SDC methods for tabular data, such as cell-key adjustment, will not be appropriate. Nonetheless, the use of other dynamic SDC tools can be an interesting challenge for future research, which can focus, e.g. on reducing the computational overload, which is unavoidable in the case of such large files.

## References

Abowd, J. M., (2018). The US Census Bureau adopts differential privacy. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 2867.

Brand, R., (2002). Microdata protection through noise addition. *Inference Control in Statistical Databases: From Theory to Practice*, pp. 97–116.

Calian, V., (2020). Methods of statistical disclosure control for aggregate data with a case study on the new Icelandic geospatial system of statistical output areas. *Working Papers of Statistics Iceland*, 105(6).

Jackson, J., Mitra, R., Francis, B., and Dove, I., (2024). Obtaining $(\varepsilon, \delta)$ - Differential Privacy Guarantees When Using a Poisson Mechanism to Synthetize Contingency Tables. *Privacy in Statistical Databases: International Conference, PSD 2024, Antibes Juan-les-Pins, France, September 25–27, 2024. Proceedings*, pp. 102–112.

Jansson, I., (2012). Issues and plans for the disclosure control of the Swedish Census 2011. *En Workshop on Statistical Disclosure Control of Census Data*, Luxembourg.

Kim, J. J., (1986). A method for limiting disclosure in microdata based on random noise and transformation. *Proceedings of the Section on Survey Research Methods*, pp. 303–308.

Kowarik, A., (2015). *New computational tools and methods for official statistics* [Doctoral dissertation, Technische Universitat Wien].

Kraus, J., (2021). Statistical Disclosure Control methods for Harmonised Protection of Census Data: A Grid Case. *Demografie*, 63(4), pp. 199–215.

Młodak, A., Pietrzak, M., and Jozefowski, T., (2022). The trade-off between the risk of disclosure and data utility in SDC: A case of data from a survey of accidents at work. *Statistical Journal of the IAOS*, 38(4), pp. 1503–1511.

Młodak, A., (2020). Information loss resulting from Statistical Disclosure Control of output data [(in Polish)]. *Wiadomości Statystyczne. The Polish Statistician*, 65(09), pp. 7–27.

Muralidhar, K., Domingo-Ferrer, J., (2023). A Rejoinder to Garfinkel (2023) – Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option. *Journal of Official Statistics*, 39(3), pp. 411–420.

Shlomo, N., (2022). How to Measure Disclosure Risk in Microdata? *The Survey Statistician*, 86, pp. 13–21.

Shlomo, N., Skinner, C., (2022). Measuring risk of re-identification in microdata: State-of-the art and new directions. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 185(4), pp. 1644–1662.

Shlomo, N., Tudor, C., and Groom, P., (2010). Data swapping for protecting census tables. *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010, Corfu, Greece, September 22–24, Proceedings*, pp. 41–51.

SNI et al., (2022). *Census 2021 Statistical Disclosure Control Methodology*. Northern Ireland Statistics & Research Agency.

Templ, M., (2017). *Statistical Disclosure Control for Microdata. Methods and Applications in R*. Springer International Publishing AG, Cham, Switzerland.

Templ, M., Kowarik, A., and Meindl, B., (2015). Statistical Disclosure Control for Micro-Data Using the R Package `sdcMicro`. *Journal of Statistical Software*, 67(4), pp. 1–36.

Tran, T., Reimherr, M., and Slavkovic, A., (2024). Differentially private quantile regression. *Privacy in Statistical Databases: International Conference, PSD 2024, Antibes Juan-les-Pins, France, September 25–27, Proceedings*, pp. 18–34.

Yoon, J., Drumright, L. N., and Van Der Schaar, M., (2020). Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8), pp. 2378–2388.

Zayatz, L., (2002). SDC in the 2000 US Decennial Census. In *Inference Control in Statistical Databases: From Theory to Practice*, pp. 193–202. Springer.