UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

**Expert Meeting on Statistical Data Confidentiality**

1-3 December 2021, Poland

# The trade-off between the risk of disclosure and data utility in SDC – a case of data from asurvey of accidents at work.

Andrzej Młodak (Statistical Office in Poznań)

*a.mlodak@stat.gov.pl*

***Abstract***

One of the key problems associated with Statistical Disclosure Control is to ensure the optimal trade-off between measures to minimize the risk of unit identification and the desire to maximize the utility of data to be disclosed (which is equivalent to minimizing information loss due to the application of SDC methods). Moreover, variables derived from statistical surveys vary not in terms of their measurement scale but also as regards the role they play in the SDC process. All these aspects should therefore be taken into account when one tries to find this trade-off.

In the paper we present a way of assessing whether an optimal trade-off has been achieved. In our study we used data from an annual survey of accidents at work for 2017. We compared complex information loss and the risk of disclosure in the original data files and those subject-ed to SDC using methods implemented in the new working version of the sdcMicro package in the R environment. In the paper we present the underlying assumptions and results of the SDC process, highlighting the benefits and drawbacks of the tools used in the study, which was conducted in 2020 and 2021 in the Centre for Small Area Estimation of the Statistical Office in Poznań.

# The trade-off between the risk of disclosure and data utility in SDC – a case of data from a survey of accidents at work

Andrzej Młodak*, Michał Pietrzak**, Tomasz Józefowski***

* Statistical Office in Poznań, Centre for Small Area Estimation; Statistical Office in Poznań, Branch in Kalisz, ul. Piwonicka 7–9, 62–800 Kalisz, Poland, e-mail: a.mlodak@stat.gov.pl; Calisia University – Kalisz, Poland, Inter-faculty Department of Mathematics and Statistics, ul. Nowy Świat 4, 62-800 Kalisz, Poland

** Statistical Office in Poznań, Centre for Small Area Estimation, ul. Wojska Polskiego 27–29, 60–624 Poznań, Poland, e-mail: m.pietrzak@stat.gov.pl

*** Statistical Office in Poznań, Centre for Small Area Estimation, ul. Wojska Polskiego 27–29, 60–624 Poznań, Poland, e-mail: t.jozefowski@stat.gov.pl

**Abstract**. One of the key problems associated with Statistical Disclosure Control is ensuring an optimal trade-off between measures aimed at minimizing the risk of unit identification and the desire to maximize the utility of data to be disseminated (which is equivalent to minimizing information loss due to the application of SDC methods). Moreover, variables from statistical surveys vary not only in terms of their measurement scale but also as regards the role they play in the SDC process. All these aspects should therefore be taken into account when one tries to find this trade-off.

In the paper we present a way of assessing whether an optimal trade-off has been achieved. Two main aspects of measuring the risk of disclosure are discussed. The first one is internal risk, i.e. the risk of disclosing confidential information only on the basis on disseminated microdata after the application of SDC (i.e. no attempt of combining data with external information is made); the second one is external risk, when the 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 that it is possible to measure external risk and information loss while accounting for the measurement scale of variables. In our empirical study we used data from an annual survey of accidents at work for 2017. We compared complex information loss and the risk of disclosure in the original data files and those subjected to SDC using methods implemented in the new working version of the sdcMicro R package. We present the underlying assumptions and results of the SDC process, highlighting the benefits and drawbacks of the tools used in the study, which was

conducted in 2020 and 2021 in the Centre for Small Area Estimation at the Statistical Office in Poznań.

# 1 Introduction

The growing demand for large microdata sets containing relevant information calls for the development and implementation of increasingly advanced methods of Statistical Disclosure Control (SDC). In recent years various new tools of SDC have been created, including those representing perturbative methods. Their optimal parameterization in practical applications is also the subject of intensive research with the help of efficient software, e.g. dedicated R packages.

However, for SDC methods to be used properly one needs to have reliable information about the effectiveness of SDC treatment, which can be assessed by using measures of the expected risk of disclosure and information loss due to the application of these methods. The expected risk of disclosure indicates the degree of ensured protection of privacy in a disclosed data set while information loss refers to the usefulness of disseminated data for analytical purposes and the degree to which they can be used to obtain reliable and pertinent results. The main goal of SDC is, therefore, to simultaneously minimize the risk of disclosure and information loss. The measures of risk should account for internal risk (identification based only on data obtained by the user without linking them with information from other sources) and external risk (identification achieved by linking records). In addition, the assessment of risk and information loss should also account for measurement scales of variables.

In this paper we discuss these aspects of disclosure risk and information loss. Internal risk is quite widely addressed in the literature, whereas external risk is considered relatively rarely (a probabilistic approach in this respect is suggested e.g. by Domingo-Ferrer and Torra (2003)). We present some tools designed to measure these aspects, including those that account for the measurement scale of particular variables, and discuss their use, advantages and drawbacks. Some of the presented measures of information loss (concerning distribution disturbance and the impact on the intensity of connection) have been implemented in the new version of the sdcMicro R package. Their usefulness and application in the SDC process is demonstrated using microdata from the Polish survey of accidents at work.

The paper is organized as follows. In Section 2 we analyze the problem of measuring the risk of disclosure (overall, internal and external) showing how external risk can be assessed for all variables, regardless of their measurement scale. Section 3 presents complex measures of information loss of various type. Section 4 contains a description of assumptions underlying the Polish survey of accidents at work, suggested flow of the SDC process in this case and results of assessing the risk of disclosure and imputation loss. The most important conclusions are summarised

in Section 5.

## 2 Assessment of disclosure risk

The assessment of the risk of identification using disclosed data involves identifying unique combinations for categorical variables or their values in the neighbourhood of relevant original values for continuous variables) or levels (individual, global or hierarchical).

One can distinguish two types of disclosure risk for a dataset obtained once the SDC process has been applied:

- internal risk – when there is a threat of identifying units only using modified data[1],
- external risk – when there is a threat of identifying units by attempting to link data after SDC with information from other sources possibly available to the user.

Internal risk results from the existence of unique combinations of values (exact for categorical variables and – if possible – within a certain precision level for continuous variables). External risk depends on the possibility of linking records contained in a statistical dataset (which underwent SDC) with relevant records from other data sources available to the user. Thus, the total risk of disclosure, which should be assessed by the statistician preparing the data for disclosure can be given as

$$r = \frac{r_{\text{int}} + r_{\text{ext}}}{2}, \tag{1}$$

where $r_{\text{int}}$ is internal risk and $r_{\text{ext}}$ is external risk. In formula (1) it is tacitly assumed that both $r_{\text{int}}$ and $r_{\text{ext}}$ take values from the same interval. It is Usually [0,1]. Moreover, it is a general idea: internal risk can be e.g. a function of 'partial' risks for particular groups of variables (categorical or continuous ones) or can express various aspects (resulting from using various principles).

Internal risk refers to the possibility of identifying a unit only based on disclosed data; of course, one can assess the risk of disclosure for original data (as a preliminary step in the SDC process), but knowing the risk associated with the data set ultimately provided to the user is more important.

The assessment of internal risk is traditionally made using well-known rules ($k$-anonymity, $l$-diversity, $t$-closeness, $(n,k)$–dominance, $p\%$, etc. - cf. e.g. Hundepool et al. (2012)) and involves assessing:

- individual risk – computed e.g. using the Benedetti – Franconi superpopulation model,

[1]It is worth noting that the measures of internal risk can obviously be used to assess the risk of disclosure in the original data.

- global risk – which is the sum of individual risks or is estimated using log–linear models or the benchmark approach.

The above mentioned measures are applied mainly to categorical variables. In the case of continuous variables, risk can be measured using the upper bound of the percentage of observations falling within an interval centered on the masked value; it can be used if perturbative SDC methods were applied (cf. Templ (2017) or Templ, Kowarik and Meindl (2015)).

The idea of external risk assumes that the user has access to an alternative data source with some (or, in the worst case, all) variables contained in the file that underwent SDC. In such cases, it is necessary to assess the risk of correctly linking records from the latter file with those in the former one. First attempts in this respect were taken by Domingo-Ferrer and Torra (2003). Their proposal is based on clusterings of two alternative data sets and probabilistic linkage. In this paper we propose another approach, taking into account the probability of the user having access to alternative knowledge about variables contained in the disclosed data set and their measurement scales.

Let us assume theoretically that an alternative data set can contain any variable found in the statistical data set. Let $x_{ij}$ be the value of the $j$–th variable ($X_j$) for the $i$–th respondent contained in the disclosed file and $x_{ij}^{\#}$ be the relevant value contained in the alternative source. External risk is assessed on the basis of the distance between records $i$ from the alternative source and $h$ in the statistical source after the application of SDC:

$$d_{ih} = \sum_{j=1}^{m} d(x_{ij}^{\#}, x_{hj}) p_j, \tag{2}$$

where $p_j \in [0,1]$ is the probability that $X_j$ will be in the alternative set, $j = 1, 2 \ldots, m$, $i = 1, 2, \ldots, n^{\#}$, $h = 1, 2, \ldots, n$. The probability $p_j$ can be assessed on the basis of the knowledge possessed by statisticians about the user who is granted access to data and after identifying other possibly accessible data sources which the users could have access to. This assessment could be supported by expert opinions. Of course, if we rule out the possibility that the user has access to variable $X_j$, we put $p_j := 0$ and if we are absolutely sure that the user has access to it, then $p_j := 1$.

The distance $d(x_{ij}^{\#}, x_{hj})$ in (2) is computed differently, depending on the measurement scale of $X_j$. If $X_j$ is nominal, then (NA is treated as a separate level)

$$d(x_{ij}^{\#}, x_{hj}) = \begin{cases} 1 & \text{if } x_{ij}^{\#} = x_{hj}, \\ 0 & \text{if } x_{ij}^{\#} \neq x_{hj}. \end{cases} \tag{3}$$

If $X_j$ is ordinal, then (NA is treated as a separate, lowest category)

$$d(x_{ij}^{\#}, x_{hj}) = \frac{\mathfrak{r}(x_{ij}^{\#}, x_{hj})}{\mathfrak{r}_j - 1}, \tag{4}$$

where $\mathfrak{r}(x_{ij}^{\#}, x_{hj})$ is the absolute difference in categories between $x_{ij}^{\#}$ and $x_{hj}$ and $\mathfrak{r}_j$ is the total number of categories of $X_j$.

If $X_j$ is continuous (i.e. it is expressed on the interval or ratio scale), then a threshold $d^* > 0$ of tolerance for closeness is established. Hence (if $x_{hj} = \text{NA}$ then $|x_{ij}^{\#} - x_{hj}| := \min_{l=1,2,\ldots,n, x_{lj} \neq \text{NA}} |x_{ij}^{\#} - x_{lj}|$ and $h := \arg\min_{l=1,2,\ldots,n, x_{lj} \neq \text{NA}} |x_{ij}^{\#} - x_{lj}|$)

$$d(x_{ij}^{\#}, x_{hj}) = \begin{cases} 1 & \text{if } \frac{|x_{ij}^{\#} - x_{hj}|}{x_{hj}} > d^*, \\ 0 & \text{if } \frac{|x_{ij}^{\#} - x_{hj}|}{x_{hj}} \leq d^*. \end{cases}$$

Records $i$ and $h$ are paired (which can be denoted as $i \bowtie h$) if $d_{ih} = 0$. Let

$$c_i = \begin{cases} 1 & \text{if } \exists_{h \in \{1,2,\ldots,n\}} i \bowtie h, \\ 0 & \text{otherwise.} \end{cases}$$

The final measure of external risk is given as

$$r_{\text{ext}} = \frac{1}{n^{\#}} \sum_{i=1}^{n^{\#}} c_i \in [0, 1].$$

An alternative data source can be simulated using the original one and assuming that data for some variables are available to the user in the alternative source. It is a radical assumption but one that makes it possible to account for all threats concerning unit identification.

## 3 Measurement of information loss

The application of SDC methods results in the loss of some information (resulting e.g. from gaps, when non-perturbative methods are used, or perturbations, when perturbative tools are used). Because of this loss the analytical worth of the disclosed data for the user decreases, which means there is a possibility that results of computations and analyses based on such data will be inadequate (e.g. the precision of estimation could be much worse).

Users should always obtain reliable information about the expected information loss (in the form of a global indicator for the whole disclosed data set and measures indicating how losses in particular variables contribute to the overall loss) in a manner which is easily understandable and interpretable. The measure of information loss is based on distances (especially normalized) between relevant values (simple values of variables, descriptive statistics of their distributions or measures of dependence or correlation) before and after the application of SDC, taking into consideration the measurement scales of particular variables.

Three main types of measures of information loss owing to SDC are traditionally distinguished:

- **measures of distribution disturbance** – measures based on distances between original and perturbed values of variables (e.g. mean, mean of relative distances, complex distances, etc.),
- **measures of impact on variance of estimation** – computed using distances between variances for averages of continuous variables before and after SDC or multi-factor ANOVA for a selected dependent variable in relation to selected independent categorical variables (in this case, the measure of information loss involves a comparison of components of coefficients of determination $R^2$ – in terms of within-group and inter-group variance – for relevant models based on original and perturbed values (cf. Hundepool et al. (2012)),
- **measures of impact on the intensity of connections** – comparisons of measures of direction and intensity of connections between original continuous variables and between relevant perturbed ones; such measures can be e.g. correlation coefficients or test of independence.

Now we present some measures of information loss, which were originally used in our study, introduced by Młodak (2019, 2020). The complex measure of distribution disturbance is given by

$$\lambda = \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{d(x_{ij}, x_{ij}^*)}{mn} \in [0,1], \tag{5}$$

where $d(\cdot,\cdot) \in [0,1]$ is measure of distance, $x_{ij}^*$ is value of $X_j$ for $i$-th unit after SDC, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots .m$.

- if $X_j$ is nominal or ordinal, then $d(\cdot,\cdot)$ is defined as in the case of the measure of external risk (formulas (3) and (4), respectively),
- if $X_j$ is continuous, then

$$d(x_{ij}, x_{ij}^*) = \frac{2}{\pi} \arctan |x_{ij} - x_{ij}^*|.$$

One can also measure the contribution of particular variables $X_j$ to total information loss as follows

$$\lambda_j = \sum_{i=1}^{n} \frac{d(x_{ij}, x_{ij}^*)}{n} \in [0,1] \tag{6}$$

If $X_j$ is nominal, then if $x_{ij}^*$ is hidden, $d(x_{ij}, x_{ij}^*) = 1$; if $X_j$ is ordinal, then we assign $x_{ij}^* := 1$ if $x_{ij}$ is closer to $k_j$ or $x_{ij}^* := k_j$ if $X_j$ is closer to 1; if $X_j$ is continuous, then

$$x_{ij}^* := \begin{cases} \max\limits_{h=1,2,\ldots,n} x_{hj} & \text{if} \quad x_{ij} \leq \text{med}_{h=1,2,\ldots,n}\, x_{hj}, \\ \min\limits_{h=1,2,\ldots,n} x_{hj} & \text{if} \quad x_{ij} > \text{med}_{h=1,2,\ldots,n}\, x_{hj}. \end{cases}$$

The measure (5) can be expressed as a percentage and shows total information loss – the greater the value of $\lambda$, the bigger the loss. In this way users obtain clear and easily understandable information about expected information loss owing to the application of SDC.

The second measure is used to assess the impact on the intensity of connections between variables. It can be applied to continuous variables. Let $\mathbf{R}$ be a correlation matrix of continuous variables in the original data set and $\mathbf{R}^*$ – the set after SDC and $\mathbf{R}^{-1}$ and $\mathbf{R}^{*-\mathbf{1}}$ – their inverses, respectively. The measure of information loss is based on diagonal entries of $\mathbf{R}^{-1}$ ($\rho_{jj}^{(-1)}$) and $\mathbf{R}^{*-\mathbf{1}}$ ($\rho_{jj}^{*(-1)}$), $j = 1, 2, \ldots, m_c$ (where $m_c$ is the number of continuous variables):

$$\gamma = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m_c} \left( \frac{\rho_{jj}^{(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho_{ll}^{(-1)}\right)^2}} - \frac{\rho_{jj}^{*(-1)}}{\sqrt{\sum_{l=1}^{m} \left(\rho_{ll}^{*(-1)}\right)^2}} \right)^2} \in [0, 1]. \tag{7}$$

Values of (7) are also easily interpretable, as the expected loss of information about connections between variables. Of course, both matrices must be based on the same correlation coefficient. The most obvious choice in this respect is the Pearson's index. However, when tau-Kendall correlation matrix is used, one can also apply it to ordinal variables. The method will be not applicable if the correlation matrix is singular.

Measures (5), (6) and (7) have been implemented in the current version of the `sdcMicro` package (cf. Templ (2021))[2].

## 4 The Polish survey of accidents at work – empirical study

The survey of accidents at work is conducted by Statistics Poland. The survey is actually administered by the Centre for Working Conditions Statistics of the Statistical Office in Gdańsk. It is an exhaustive, permanent survey carried out annually and covering all accidents at work or equivalent. Data in the survey come from the statistical accident form (Z–KW form), which has to be completed after any accident at work or equivalent according to current regulations; the form contains

[2]In the paper by Młodak (2020) the coefficient (7) has a constant multiplier equal to 1/2. This is because the formula of this index is based on the Euclidean distance between two points belonging to the unit ball on $\mathbb{R}^{m_c}$ (i.e. a sphere with the centre at origin – $(0, 0, \ldots, 0)$ – and radius 1). Of course, the maximum possible distance of that form in this case amounts to 2. Therefore, to normalize the index on [0,1] the distance should be multiplied by 1/2. However, the diagonal entries of the inverse correlation matrix are not smaller than 1. Therefore, the distance is actually computed only within the hyperquadrant of the unit circle only including points with non-negative coordinates. The maximum distance between two points, given this restriction, is equal to $\sqrt{2}$. As a result, the values of the index in the original form tended to be too small. This was the reason why in `sdcMicro` 1/2 was replaced by $1/\sqrt{2}$.

information about persons injured in the accident, about the accident itself and about the company where the accident happened.

The database to be disseminated contains about 88,000 records and 34 variables. 28 variables were identified as quasi-identifiers and 6 as non-confidential ones. Following suggestions expressed by Templ (2017), the quasi–identifiers were divided into:

- key categorical quasi-identifiers (9),
- other categorical quasi-identifiers for which PRAM was used (13),
- continuous quasi-identifiers (6).

The key categorical quasi-identifiers include most sensitive categorical variables describing:

- the number of persons employed by the company (LPB1),
- the victim's sex (P1),
- the victim's age group (P2),
- the victim's citizenship (P3),
- the victim's injury category (P8),
- the accident's geographical location (P15),
- the place category of the accident (P20)
- the company's NACE code (PKD).

The province – i.e. Polish NUTS 2 region – where the company is located (WOJW) was established as a ghost variable based on P15. It is for these variables (except for PKD[3]) that internal and external risk of disclosure is computed (none of the remaining variables was considered to be a key one).

The global risk was also computed separately for continuous variables: seniority (P6), the number of worked hours (P7), worktime lost by other employees (P13), estimated material losses caused by the accident (P14), the time of the accident (P17) and the number of months elapsed since 01/2000 to the date of the accident (data_d).

The PRAM variables included employment status (P4), the victim's occupation (P5), the body part injured (P9), the place where the accident occurred (P18), P21 – the activity performed by the victim at the time of the accident (P21), the material factor related to the activity performed by the victim at the time of the accident

[3]Due to a large number of categories for the PKD variable, targeted record swapping was applied. Moreover, the PKD variable was not taken into consideration when internal risk was computed. It was an original solution inspired by suggestions made by Templ (2017).

(P22), an event that is a deviation from the normal state (P23), material factor associated with the deviation (P24), the trauma event (25), the material factor that was the source of the injury (P26), the main cause of the accident (P27_1), consequences of the accident (P28), and the number of days of incapacity for work (P29, class intervals).

All the computations were performed in R (using the `sdcMicro` and `recordSwapping` packages).

Two variants of SDC were used, including the following common steps:

- **Targeted Record Swapping (TRS)** with the following parameters: hierarchy – PKD (NACE division and group), hid – specially created ID of the accident, similar – LPB1 and SEK (sector of ownership), swaprate – 0.05, k_anonymity – $k = 3$, risk_variables – P1, P2, P3 and P28 and seed – 123; it is worth noting that this method is usually used for hierarchies of spatial units, but in this case it was applied for domain hierarchy,
- **Post-RAndomization Method (PRAM)** for PRAM variables with parameters pd=0.70 and alpha=0.30,
- **Noise addition** for continuous variables using method ="correlated2" and delta = 0.75[4].

The following variants differed in their treatment of key categorical quasi-identifiers other than PKD and WOJW:

- **variant I**: local suppression (LS) was applied; for the $k$-anonymity rule: $k = 3$; the importance of particular variables was established on the basis of their contribution to the global risk computed using the SUDA approach (cf. Templ (2017)) – the vector of contributions was of the form c(LPB1,P1,P2,P3,P8,P15,P20) = (4,1,4,2,7,1,6), and the order of considered combinations of particular sizes was set as combs=(3,4,5,6,7),
- **variant II**: micro-aggregation based on the Gower distance (MGD) was used; the clustering was performed on the basis of variables LPB1, P1, P2, P3, P8, P20; the distance between records was computed using dist_var – P1, P2, P3; minimal expected cluster size aggr – 4; micro-aggregation was performed within the province of the place of the accident, i.e. by – P15, and the micro-aggregation was performed using the maxCat function (which chooses as perturbation the level with the most occurrences in a given group in the original data set or is set to random if the maximum is not unique).

First, the number and percentages of combinations violating the $k$-anonymity rule were computed. In variant I we obtained:

[4]This option has given good results; however, in practice lower values are recommended, e.g.0.10–0.15

- 2–anonymity: 0 (0.000%); for original data: 3189 (3.610%),
- 3–anonymity: 0 (0.000%); for original data: 6079 (6.882%),
- 5–anonymity: 3289 (3.724%); for original data: 11354 (12.854%)

and in variant II:

- 2–anonymity: 119 (0.135%); for original data: 3189 (3.610%),
- 3–anonymity: 221 (0.250%); for original data: 6079 (6.882%),
- 5–anonymity: 427 (0.483%); for original data: 11354 (12.854%).

In variant I the ideal satisfaction of 2- and 3-anonymity resulted from the properties of this method. However, in variant II the percentages of records violating the $k$-anonymity rules are very small. What's more, the reduction in the risk of 5-anonymity violation is much smaller for variant II than for variant I. Tables 1 and 2 show the basic descriptive statistics for global individual risk and indicators of global risk, computed using the `sdcMicro` package (the Benedetti-Franconi model).

| Statistics | Values | |
|---|---|---|
| | original | after LS |
| **Individual risk** | | |
| Minimum | 0.005587 | 0.005587 |
| First quartile | 0.250000 | 0.250000 |
| Median | 0.500000 | 0.500000 |
| Mean | 0.622823 | 0.593017 |
| Third quartile | 1.000000 | 1.000000 |
| Maximum | 1.000000 | 1.000000 |
| **Global** | | |
| Risk in % | 62.28235 | 5.967048 |
| Expected number of re-identifications | 55014 | 5270.694 |
| Threshold | 0.008547 | 0.010417 |

Table 1: Basic descriptive statistics of individual risk for key variables – variant I (using Local Suppression - LS)

Source: Results produced using the `sdcMicro` R package.

Variant II ensures a more significant reduction of risk than Variant I. The maximum risk for continuous variables amounted to 0.00% in both cases, which indicates that noise addition is efficient (even if delta ranges from 0.10 to 0.15, which was verified).

| Statistics | Values | |
|---|---|---|
| | original | after MGD |
| **Individual risk** | | |
| Minimum | 0.005587 | 0.003876 |
| First quartile | 0.250000 | 0.058824 |
| Median | 0.500000 | 0.166667 |
| Mean | 0.622823 | 0.300023 |
| Third quartile | 1.000000 | 0.500000 |
| Maximum | 1.000000 | 1.000000 |
| **Global risk** | | |
| Risk in % | 62.28235 | 0.9158836 |
| Expected number of re-identifications | 55014 | 809 |
| Threshold | 0.008547 | NA |

Table 2: Basic descriptive statistics of individual risk for key variables – variant II (using MIcroaggregation based on the Gower distance - MGD)

Source: Results produced using the `sdcMicro` R package.

In the simulated computation of external risk according to formula (2) for any key quasi–identifier it was rigorously assumed that $p_j = 1$ and $p_j = 0$ for the remaining variables, i.e. we are sure that the user has full information from other sources about all key quasi–identifiers and no additional information about remaining variables. Moreover, we assume that the data are up-to-date and have no gaps or other errors. For better efficiency, in the algorithm a given record from the original data set is linked with the first exactly adjusted record found in the file after SDC. In other words, the order of records is also important.

In effect, for Variant I the key quasi–identifiers enabled us to correctly identify as many as 63,897 records (72.33896%) and for Variant II – only 6,998 records (7.922563%). Thus, Variant II is more effective and produces a safer output file.

However, the main problem connected with this algorithm was that the computation involves a comparison of $n(n+1)/2$ pairs of records, where $n$ is the total number of records (if a record from the file after SDC is correctly linked with a record from the original file based on key categorical quasi-identifiers, then it is not taken into account in the next round), which is seriously time-consuming (in our case it took about 8 days). Hence, further time reduction seems to be a challenge.

Tables 3 and 4 present results regarding information loss for Variant I and Variant II computed using formula (6)

As can be seen, information loss is similar for both variants. The biggest differences (to the disadvantage of variant II) occur in the case of variables: LPB1, P8 and P20. It can be due to larger clusters in micro-aggregation and the fact that there is little variation between values of these variables within these clusters. Similar levels of information loss were also observed in the case of total information loss (20.8% in variant I and 24.7% in variant II) as well as for information loss regarding correlation between continuous variables (2.8% and 3.3%, respectively).

| Variable | Loss | Variable | Loss |
|---|---|---|---|
| WOJW | 0.004 | P17 | 0.850 |
| LPB1 | 0.001 | P18 | 0.072 |
| P1 | 0.000 | P19 | 0.000 |
| P2 | 0.000 | P20 | 0.000 |
| P3 | 0.000 | P21 | 0.086 |
| P4 | 0.029 | P22 | 0.071 |
| P5 | 0.078 | P23 | 0.067 |
| P6 | 0.973 | P24 | 0.059 |
| P7 | 0.878 | P25 | 0.071 |
| P8 | 0.038 | P26 | 0.070 |
| P9 | 0.066 | P27_1 | 0.091 |
| P10 | 0.000 | P28 | 0.006 |
| P13 | 0.952 | P29 | 0.043 |
| P14 | 0.996 | PKD | 0.071 |
| P15 | 0.004 | data_d | 0.665 |

Table 3: Information loss on particular variables – variant I

Source: Own elaboration using `sdcMicro` package of the R environment.

| Variable | Loss | Variable | Loss |
|---|---|---|---|
| WOJW | 0.000 | P17 | 0.876 |
| LPB1 | 0.252 | P18 | 0.050 |
| P1 | 0.000 | P19 | 0.000 |
| P2 | 0.000 | P20 | 0.308 |
| P3 | 0.000 | P21 | 0.078 |
| P4 | 0.056 | P22 | 0.059 |
| P5 | 0.067 | P23 | 0.073 |
| P6 | 0.963 | P24 | 0.061 |
| P7 | 0.867 | P25 | 0.064 |
| P8 | 0.642 | P26 | 0.076 |
| P9 | 0.100 | P27_1 | 0.075 |
| P10 | 0.000 | P28 | 0.006 |
| P13 | 0.955 | P29 | 0.036 |
| P14 | 0.997 | PKD | 0.071 |
| P15 | 0.000 | data_d | 0.665 |

Table 4: Information loss on particular variables – variant II

Source: Own elaboration using `sdcMicro` package of the R environment.

## 5 Conclusions

Several important conclusions can be drawn from the study presented above. To achieve the right balance between the risk of disclosure and information loss one

should account for all important measurable aspects. Such an approach would guarantee reliable and exhaustive information about these key aspects of SDC.

Total risk of disclosure involves threats of re-identification based either exclusively on information from a statistical data set or achieved by linking disclosed data with relevant records from other sources available to the user; in the latter case one should estimate the probability of the user getting access to such data – based on information about the user and previous experience. One problem associated with the measure of external risk is relatively long computation time – especially for larger data sets and assumptions used in the attempt to assess such risk.

The measures of external risk and information loss that account for the measurement scales of variables can provide reliable and pertinent information about these problems; information that can be clearly interpreted by statistician (risk of disclose and information loss) and users (information loss).

As shown in our study (and, arguably, in many other cases) the application of perturbative SDC methods seems to provide better effects than the use of non-perturbative ones. While the extent of information loss is similar for both methods, perturbation reduces the risk of disclosure risk more effectively.

## References


Domingo-Ferrer, J. and Torra, V. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing, **13**(4), 343–354.

Hundepool, A., Domingo–Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P.–P. (2012), *Statistical Disclosure Control*, series: Wiley Series in Survey Methodology, John Wiley & Sons, Ltd., Chichester.

Młodak, A. (2020). Information loss resulting from statistical disclosure control of output data, *Wiadomości Statystyczne. The Polish Statistician* **65**(9):7–27, DOI: 10.5604/01.3001.0014.4121 (in Polish)

Młodak, A. (2019). Using the Complex Measure in an Assessment of the Information Loss Due to the Microdata Disclosure Control, *Przegląd Statystyczny. Statistical Review*, **66**(1):7–26, DOI: 10.5604/01.3001.0013.8285 (in Polish).

Templ, M. (2021), Package 'sdcMicro', version 5.6.1., July 26, 2021, `https://cran.r-project.org/web/packages/sdcMicro/sdcMicro.pdf`.

Templ, M. (2017), *Statistical Disclosure Control for Microdata Using 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):1–36.