

*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, Poland

Poznań, 1st December 2021

Introduction

- The growing demand for large microdata sets containing relevant information requires the development and use of increasingly advanced methods of Statistical Disclosure Control (SDC).
- Efficient use of SDC methods depends on reliable estimates of disclosure risk and information loss caused by their application.
- The main goal of SDC is to simultaneously minimize the risk of disclosure and the loss of information.
- Measures of risk should account for internal risk (identification based exclusively on data contained in a disclosed set) and external risk (identification by linking records with data from alternative data sets that a user may have access to).
- When assessing disclosure risk and information loss, one should also consider the measurement scale of variables.
- In the presentation we'll show some tools to measure these aspects and how they were applied to the Polish survey of accidents at work.

Presentation scheme

- 1 Assessment of disclosure risk
- 2 Measurement of information loss
- 3 Polish survey of accidents at work
- 4 Conclusions and references

Assessment of disclosure risk

What is the risk of disclosure?

- The risk of identification using disclosed data is assessed by identifying unique combinations of values (exact for categorical variables and within a certain precision level for continuous variables) or levels of risk (individual, global or hierarchical).
- Types of disclosure risk for a modified dataset (after applying SDC):
 - internal risk – when there is a threat of identifying units only using modified data,
 - external risk – when there is a threat of identifying units by attempting to link modified data with information from other sources possibly available to the user.
- Thus, the total risk of disclosure to be assessed before data are disclosed can be given as

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

where r_{int} is internal risk and r_{ext} is external risk.

Assessment of disclosure risk

Internal risk

- Internal risk refers to the possibility of identifying a unit only based on disclosed data; of course, the risk of disclosure can be assessed for original data, but it is more important to know the risk for the data set ultimately provided to the user.
- Internal risk is traditionally assessed using well-known rules (k -anonymity, l -diversity, t -closeness, (n, k) -dominance, $p\%$, etc.) and involves assessing:
 - individual risk – computed e.g. using Benedetti-Franconi super-population model,
 - global risk – the sum of individual risks or estimated using log-linear models or the benchmark approach.
- These measures are applied mainly to categorical variables. Risk for continuous variables 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)).

Assessment of disclosure risk

External risk

- The user is assumed to have access to an alternative data source with some (or, in the worst case, all) variables contained in the file that underwent SDC.
- It is necessary to assess the risk of correctly linking records from the latter file with those from the former one.
- Let $x_{ij}^{\#}$ be the value of the j -th variable (X_j) for the i -th respondent contained in the alternative source. External risk is assessed on the basis on the distance between records i in the alternative source and h in the statistical source after the application of SDC:

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

where $p_j \in [0, 1]$ is the probability that X_j will be in the alternative set, $j = 1, 2, \dots, m$, $i = 1, 2, \dots, n^{\#}$, $h = 1, 2, \dots, n$.

- The probability p_j can be assessed using the statistician's knowledge about the user who is given access to data and after identifying other possibly accessible data sources which the user could have access to.

Assessment of disclosure risk

External risk

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

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

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

$$d(x_{ij}^{\#}, x_{hj}) = \frac{\tau(x_{ij}^{\#}, x_{hj})}{\tau_j - 1},$$

where $\tau(x_{ij}^{\#}, x_{hj})$ is the absolute difference in categories between $x_{ij}^{\#}$ and x_{hj} and τ_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), a threshold $d^* > 0$ of tolerance for closeness is established.

Assessment of disclosure risk

External risk

- Hence (if $x_{hj} = \text{NA}$ then $|x_{ij}^{\#} - x_{hj}| := \min_{l=1,2,\dots,n, x_{lj} \neq \text{NA}} |x_{ij}^{\#} - x_{lj}|$ and $h := \arg \min_{l=1,2,\dots,n, x_{lj} \neq \text{NA}} |x_{ij}^{\#} - x_{lj}|$)

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

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

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

- The final measure of external risk is given as $r = \frac{1}{n^{\#}} \sum_{i=1}^{n^{\#}} c_i \in [0, 1]$.
- An alternative source can be simulated using the original one and assuming that data for some variables are available to the user in the alternative source.

Measurement of information loss

The role of information loss in SDC

- 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 that results of computations and analyses based on such data may be inadequate.
- 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.

Measurement of information loss

Types of measures

- **Measures of distribution disturbance** – measures based on distances between original and perturbed values of variables (e.g. mean, mean of relative distances, etc.),
- **Measures of impact on the variance of estimates** – computed using distances between variances for averages of continuous variables before and after SDC or multi-factor ANOVA for selected dependent variables 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 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 perturbed ones; such measures can include correlation coefficients or tests of independence.

Measurement of information loss

Examples of measures

- Measure of distribution disturbance

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

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

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

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

- measure λ can be expressed as a percentage and shows total information loss – the greater the value of λ , the bigger the loss.

Measurement of information loss

Examples of measures

- Measure of distribution disturbance

- One can also measure the contribution of X_j to total information loss as

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

- if X_j is nominal, then if x_{ij}^* is hidden, then $d(x_{ij}, x_{ij}^*) = 1$; 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

$$x_{ij}^* := \begin{cases} \max_{h=1,2,\dots,n} x_{hj} & \text{if } x_{ij} \leq \text{med}_{h=1,2,\dots,n} x_{hj}, \\ \min_{h=1,2,\dots,n} x_{hj} & \text{if } x_{ij} > \text{med}_{h=1,2,\dots,n} x_{hj}. \end{cases}$$

Measurement of information loss

Examples of measures

- Measure of impact on the intensity of connections
 - can be applied to continuous variables and is based on diagonal entries of an inverse correlation matrix before $(\rho_{jj}^{(-1)})$ and after SDC $(\rho_{jj}^{*(-1)})$, $j = 1, 2, \dots, m_c$ (m_c – the number of continuous variables):

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

- values of γ are easily interpretable. In the case of a tau-Kendall correlation matrix, ordinal variables can also be used. The method is not applicable if the correlation matrix is singular.
- Measures λ and γ are discussed by Młodak (2019) and Młodak (2020) (with some variations) and implemented in the current version of the `sdcmicro` package (cf. Templ (2021)).

Polish survey of accidents at work

Basic information

- Conducted by the Centre for Working Conditions Statistics of the Statistical Office in Gdańsk.
- It is an exhaustive, permanent survey covering all accidents at work or equivalent, its results are published annually.
- Survey data come from the statistical accident form completed after any accident at work or equivalent according to current regulations; it contains 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 of 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).

Polish survey of accidents at work

Assumptions of the SDC process

- 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 – Polish NUTS 2 region – where the company is located (WOJW – ghost variable based on P15); internal and external risk of disclosure is computed for these variables
- 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 exact time of the accident (P17), the number of months elapsed since 01/2000 to the date of the accident (data_d)
- All the computations were performed in R (using the `sdcMicro` and `recordSwapping` packages).

Polish survey of accidents at work

SDC methods used in the study

- Two variants of the SDC process were used, including the following common steps:
 - **Targeted Record Swapping**, TRS: hierarchy – PKD (NACE division and group), hid – specially created ID of accident, similar – LPB1 and SEK (sector of ownership), swaprata – 0.05, k_anonymity – $k = 3$, risk_variables – P1, P2, P3 and P28, seed – 123,
 - **Post-Randomization Method**, PRAM: pd=0.70, alpha=0.30,
 - **Noise addition** for continuous variables: method="correlated2", delta = 0.75.
- The two variants differed in their treatment of key categorical quasi-identifiers other than PKD and WOJW:
 - **variant I**: local suppression, LS ($k = 3$, contribution of particular variables to the global risk computed using the SUDA approach – $w(LP1,P1,P2,P3,P8,P15,P20)=(4,1,4,2,7,1,6)$, combs – 3,4,5,6,7),
 - **variant II**: microaggregation based on Gower's distance, MG (variables – LPB1, P1, P2, P3, P8, P20, dist_var – P1, P2, P3, aggr – 4, by – P15, maxCat).

Polish survey of accidents at work

Results of the assessment of internal risk of disclosure

- Number and percentages of combinations violating the k -anonymity rule
 - variant I:
 - 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%).
 - 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%).

Polish survey of accidents at work

Results of the assessment of internal risk of disclosure

- Individual risk and global risk – variant I
 - descriptive statistics for individual risk and global risk

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

- The maximum risk for continuous variables amounted to 0.00%

Polish survey of accidents at work

Results of the assessment of internal risk of disclosure

- Individual risk and global risk – variant II
 - descriptive statistics for individual risk and global risk

Statistics	Values	
	original	after MG
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

- The maximum risk for continuous variables amounted to 0.00%

Polish survey of accidents at work

Results of the assessment of external risk of disclosure

- For any key quasi-identifier it was rigorously assumed that $p_j = 1$ and $p_j = 0$ for other variables, i.e. the user is assumed to have full information from alternative sources on all key quasi-identifiers and no additional information about remaining variables.
- Variant I: 63897 (i.e. 72.33%) records were correctly identified using the key quasi-identifiers
- Variant II: Only 6998 (i.e. 7.92%) records were correctly identified using the key quasi-identifiers.
- The main problem: the computation involves a comparison of $n(n+1)/2$ pairs of records, which is very time-consuming (it took us about 8 days).

Polish survey of accidents at work

Measures of information loss

- Variant I

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

- total information loss – 20.8%,
- information loss regarding correlation between continuous variables – 2.8%.

Polish survey of accidents at work

Measures of information loss

- Variant II

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

- total information loss – 24.7%,
- information loss regarding correlation between continuous variables – 3.3%.

Conclusions and references

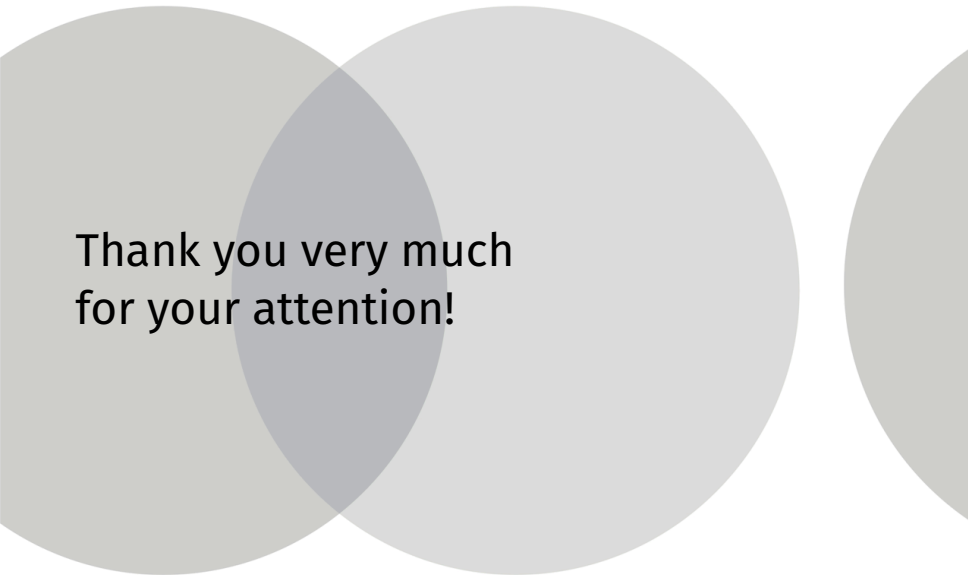
Conclusions

- To achieve the right balance between the risk of disclosure and information loss one should account for all important measurable aspects.
- 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 information about these problems; information that can be clearly interpreted by statisticians (risk of disclosure 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.

Conclusions and references

References

- Hundepool, A., Domingo–Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., 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., Meindl, B. (2015), Statistical disclosure control for microdata using the R package sdcmicro. *Journal of Statistical Software*, 67(4):1–36.



Thank you very much
for your attention!