Statistical Disclosure Control Methods for Microdata from the Labour Force Survey

Pietrzak, Michał

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Pietrzak, M. (2020). Statistical Disclosure Control Methods for Microdata from the Labour Force Survey. *Acta Universitatis Lodziensis. Folia Oeconomica*, *3*(348), 7-24. https://doi.org/10.18778/0208-6018.348.01

Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
3(348) 2020

DOI: http://dx.doi.org/10.18778/0208-6018.348.01

Michał Pietrzak
Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics Department of Statistics; Statistical Office in Poznań, m.pietrzak@stat.gov.pl

Statistical Disclosure Control Methods for Microdata from the Labour Force Survey

Abstract: The aim of this article is to analyse the possibility of applying selected perturbative masking methods of Statistical Disclosure Control to microdata, i.e. unit-level data from the Labour Force Survey. In the first step, the author assessed to what extent the confidentiality of information was protected in the original dataset. In the second step, after applying selected methods implemented in the *sdcMicro* package in the R programme, the impact of those methods on the disclosure risk, the loss of information and the quality of estimation of population quantities was assessed. The conclusion highlights some problematic aspects of the use of Statistical Disclosure Control methods which were observed during the conducted analysis.

Keywords: Statistical Disclosure Control, perturbative methods, PRAM, Additive Noise, Rank Swapping, microdata, Labour Force Survey, sdcMicro package

JEL: C18, H83, J20

1. Introduction

This article is a response to the growing demand, observed in recent years, for increasingly detailed statistical information which is of interest to both the public sector (authorities, administration, universities, or research institutes) and the private sector (business entities). Aggregated statistical tables do not fully satisfy such needs. Unit-level data (microdata) are the kind of information users really expect from statistical offices. However, microdata contain variables that directly or indirectly enable the identification of individual statistical units, so they must be duly prepared before being released. Individual units are protected against potential identification not only by Polish and European law, but also by virtue of ethical principles, which means that sensitive data cannot be disclosed. Microdata should be verified to eliminate or reduce the risk of disclosure, while simultaneously minimising the loss of information. This process is called *Statistical Disclosure Control* (SDC). Before being shared or published, all statistical information (in the form of unit-level data, statistical tables, descriptive statistics, analytical results of analyses and descriptions, charts, etc.) should be subjected to SDC. While the publication of statistical tables has a long tradition, making unit-level data available is a relatively new practice – the first attempt was made in the 1960s on data from the American census, and in 1971 Statistics Canada released *Public Use Microdata File* (PUMF) from the Canadian census (Duncan, Elliot, Salazar-González, 2011). Given the existing and growing demand for information, in order to satisfy users' expectations while protecting confidential information, SDC will become an integral part of every statistical survey.

This article documents preliminary work aimed at developing a universal solution for how to prepare datasets from sample surveys conducted by Statistics Poland so that they can be made available for scientific purposes.

The empirical study was conducted on microdata from the Labour Force Survey (LFS) – a representative, quarterly survey, conducted since 1992 to collect information about the size and structure of the labour force in Poland. Information on the survey's methodology can be found in quarterly publications (e.g. CSO, 2012).

The main SDC tool used in the study is the *sdcMicro* package for the open source R programme, created by three employees of the Austrian Statistical Office (Templ, Kowarik, Meindl, 2015), in which i.a. selected perturbative methods are implemented. The *PROC SURVEYMEANS* procedure of the SAS programme (Lewis, 2016) was used to estimate the unemployment rate and determine the estimation precision.

The aim of the article is to analyse the possibility of using selected perturbative methods to protect LFS microdata against the risk of disclosure. First, the confidentiality of the original microdata was evaluated. Then, after applying selected perturbative methods, an assessment was made of their impact on the risk

of disclosure, the loss of information and the quality of estimation (specifically the *unemployment rate*, at the province level, total and by *Sex* – as defined in the LFS, i.e. *the percentage of unemployed people in economically active population, aged from 15 to 74*).

The article is divided into six sections. The first one introduces the idea of SDC, the research problem, the objectives, and the structure of the article. The second section presents selected methods of evaluating microdata confidentiality which are available in the *sdcMicro* package. The third section is devoted to a description of selected perturbative methods for microdata. In the fourth section, attention is focused on how the problem of information loss is handled by the *sdcMicro* package as well as on point estimates and their quality. The penultimate section contains a discussion of empirical results. The article ends with conclusions, a brief summary of problematic aspects of SDC for microdata, and an outline of further research work.

2. The measurement of microdata confidentiality

In addition to the commonly known classification of variables in microdata into *continuous* and *categorical*, which has an impact on the data preparation process, the risk of disclosure and information loss, another basic division, involves four non-disjoint categories (Willenborg, de Waal, 2001; Domingo-Ferrer, Torra, 2003; Hundepool et al., 2010; 2012; Templ, Kowarik, Meindl, 2015; Templ, 2017; Benschop, Machingauta, Welch, 2019):

1) *direct identifiers* – variables that directly identify respondents;
2) *quasi-identifiers* (*key variables*) – a set of variables that, in combination, may result in re-identification of (some) respondents; potentially, each variable can be a quasi-identifier;
3) *sensitive variables* (*confidential*) – variables that contain sensitive information about respondents;
4) *non-sensitive variables* (*non-confidential*) – variables that do not contain sensitive information about respondents but may be quasi-identifiers, so they cannot be ignored in the process of ensuring the confidentiality of the dataset.

At the start of the SDC process, it is normally assumed that microdata have been anonymised (by removing identifiers), but do contain quasi-identifiers, sensitive and non-sensitive variables, which must all be subjected to SDC.

In the literature, disclosure is usually divided into the following three types (Hundepool et al., 2010; 2012; Templ, Kowarik, Meindl, 2015; Templ, 2017; Benschop, Machingauta, Welch, 2019):

1) *identity disclosure* – the intruder has managed to link a given respondent to their corresponding record in the released microdata;

2) *attribute disclosure* – the intruder has learnt some of the respondent's features from the released microdata;
3) *inferential disclosure* – the intruder is able to infer the value of some of the respondent's features more accurately from the released dataset.

SDC methods for microdata provide protection against the first two types of disclosure.

The measurement of disclosure risk varies depending on the nature of quasi-identifiers. For categorical variables, the risk is usually determined based on the uniqueness of values. For continuous variables, due to a very large or infinite number of possible values, risk assessment is based on the uniqueness of values in the neighbourhood of original values. The following approaches are presented below: the *k*-anonymity principle and the expected number of re-identification (for categorical variables) and risk determined in the interval format (for continuous variables). Both measures are available in the *sdcMicro* package.

The *k-anonymity* rule is satisfied if the number of units sharing the same combination of categorical quasi-identifier values is greater than or equal to a fixed threshold *k*. If any unit violates this rule for $k = 2$, it is considered to be sample unique. The measure of risk is the number of records in microdata that violate the *k*-anonymity rule for a fixed *k*. With regard to sample surveys, this rule does not take into account sampling weights. If they are higher, units in the sample represent more units in the population, so the probability of disclosure is lower and a lower threshold can be selected – also when the sampling weights are higher, the sample size tends to be smaller, so the number of units sharing a combination of categorical key variables is likely to be lower. Non-responses are treated as 'any other value' when measuring risk. The use of this principle alone to ensure the safety of microdata is not sufficient. For example, if all observations with the same combination of categorical key variables share the same value for another confidential variable, then there is a risk of attribute disclosure even though there is no risk of identity disclosure. For this reason, the *l-diversity* principle can be applied as a complementary rule.

The expected number of re-identification is obtained by multiplying the global risk (in percent) and the number of observations. The global risk is the sum of individual risk determined for each record in the microdata set, based on the frequency of combinations of values of categorical key variables in the sample and in the population (see: Benschop, Machingauta, Welch, 2019). The expected number of re-identifications depends on the hierarchical structure of data – the inclusion of higher-level units increases the global risk, and consequently, the expected number of re-identifications.

As mentioned above, methods of assessing the risk of disclosure – in the case of continuous variables – are based on uniqueness of values (defined in absolute or in relative terms) in the neighbourhood of original values – most of them are evaluated after anonymisation (a posteriori) (Benschop, Machingauta, Welch, 2019).

In the *sdcMicro* package, the default measure of disclosure risk for continuous variables is *interval disclosure* (Templ, Kowarik, Meindl, 2015; Templ, 2017; Benschop, Machingauta, Welch, 2019) – intervals are created around each perturbed value and then the algorithm determines whether the original value of that perturbed variable is included in the interval. If the value falls within the interval around the perturbed value, it is considered too close to the original value and deemed unsafe (requires more perturbation). If the value lies outside of the interval, it is considered safe. The size of intervals is based on the standard deviation and a scaling parameter. It is not a sufficient measure for outliers, because outliers will remain outliers (even if a perturbative method is applied) and are easily re-identifiable – even if they are sufficiently far from their initial values.

The output of the R package shows the percentage of observations falling within an interval centred on its masked value and the disclosure risk is expressed as an interval where the upper limit corresponds to the worst case scenario (in which the intruder is sure that each nearest neighbour is indeed the true link). If continuous quasi-identifiers are not anonymised, the disclosure risk can be high – up to 100%.

A review of methods for assessing the risk of disclosure can be found in: Domingo-Ferrer and Torra (2004), Hundepool et al. (2010; 2012), Shlomo (2010), Matthews and Harel (2011), Templ, Kowarik and Meindl (2015), Templ (2017), Benschop, Machingauta and Welch (2019). The issue raised in the literature in the context of risk assessment is the case of microdata with continuous and categorical quasi-identifiers. The possibility of using several approaches can be found in Hundepool et al. (2012).

3. Selected perturbative masking methods

SDC methods can be divided into two groups (Willenborg, de Waal, 2001; Hundepool et al., 2010; 2012; Shlomo, 2010; Matthews, Harel, 2011; Templ, Kowarik, Meindl, 2015; Templ, 2017; Benschop, Machingauta, Welch, 2019):

1) *non-perturbative method*s do not change original values of the variable; some values of the variable are concealed, the granularity of the variable is reduced, or only some observations are shown;
2) *perturbative methods* change variable values for some or for all observations to ensure that similar results can be obtained for the population to those that could be obtained from the original microdata; these methods replace some unique combinations of quasi-identifier values with other values, so that the intruder can never be sure whether the value of the perturbed variable is true, and consequently, whether the matching of (some) microdata records to the external database is accurate.

Each of above-presented approaches has advantages and disadvantages. Non-perturbative methods are associated with information loss. In this case, it may be necessary to apply some imputation methods or calibrate sampling weights. These negative effects do not occur when perturbative methods are used, but some inconsistencies in the dataset may occur instead, certain data patterns may be perturbed, and the perturbation applied to variables will affect the results of statistical analysis.

In the empirical study, three perturbative methods were used – *PRAM*, *Additive Noise*, and *Rank Swapping*.

PRAM (*Post-RAndomisation Method*) is a perturbative method for categorical variables. It is a probabilistic method in which values of a categorical variable are replaced with others, with a probability specified in a Markov matrix (transition matrix) – a square matrix whose rows and columns correspond to categories (levels) of the variable. The transition matrix contains probabilities – the value at the intersection of a given row and a given column is the probability of changing the category represented by the row into the category represented by the column. Probabilities in each row must add up to 100%. It is possible to exclude undesirable changes by properly constructing the Markov matrix. A special case is the *invariant PRAM*, an approach that guarantees consistency of the variable distribution before and after the application of PRAM. The Invariant PRAM guarantees that univariate tabulations remain the same, but this does not apply to cross-tabulations of variables. In the case of sample surveys, each observation may have a different sampling weight, so that, after generalisation, consistency is not assured. This method is recommended when there are at least six quasi-identifiers in microdata or when the use of non-perturbative methods would result in excessive information loss (Hundepool et al., 2012; Templ, 2017). If there are non-responses in the categorical variable, they are not replaced with any of the variable levels; likewise, values of the perturbed variable are not replaced with a non-response.

Additive Noise is a perturbative method for continuous variables. In the *Noise Addition* method, it is assumed that a perturbative vector (representing a random variable) ε_j will be added to the vector x_j of values of the *j*-th variable in the original microdata:v

$$z_j = x_j + \varepsilon_j, \tag{1}$$

where:

$$\varepsilon_j \sim N\left(0;\sigma^2_{\varepsilon_j}\right) and\ cov\left(\varepsilon_t,\varepsilon_l\right) = 0 \text{ for all } t \neq l\ .$$

It is assumed that the variance of the random variable ε_j is proportional to the variance of the original variable. This approach preserves the expected value and

covariances, while the variance and correlation coefficients are not preserved (for proof, see: Hundepool et al., 2012). Non-responses are not perturbed in the Additive Noise method (or in other Noise Addition methods).

Rank Swapping is a perturbative method for variables measured using an ordinal or higher-level scale, and is a special case of the *Data Swapping* method – based on the idea of exchanging values of the confidential variable between records. Firstly, values of the variable are ranked in ascending order and then each ranked value of the variable is swapped with another ranked value randomly chosen within a restricted range. The original version of the Rank Swapping method does not prevent attribute disclosure because it only reorders data. For example, if the intruder knows which unit has the highest value for the confidential variable, they can easily find it in the microdata (assigning this value to a different record does not help). One solution which secures the lowest and the highest values of the perturbed variable involves grouping them first and then replacing them with an average value (this approach is available in the *sdcMicro* package). Non-responses are not perturbed in the Rank Swapping method.

4. The measurement of microdata utility

The R programme report does not contain a universal measure of information loss in the case of perturbing categorical variables, although two measures are available for continuous variables – *IL*1 and *Difference of Eigenvalues.*

The first one is given by the formula:

$$IL1 = \frac{1}{p}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left|x_{ij} - x'_{ij}\right|}{\sqrt{2}S_j}, \tag{2}$$

where:
x_{ij} – the value of the i-th observation of the j-th variable in original microdata **X**,
x'_{ij} – the value of the i-th observation of the j-th variable in perturbed microdata **X'**,
$n \times p$ – dimensions of microdata,
S_j – standard deviation of the j-th original variable.

This measure is useful for comparing different methods. The smaller the value of this measure, the closer the values of the perturbed variable are to the original values, and, consequently, the higher their utility but also the risk of disclosure.

The second measure compares relative absolute differences between eigenvalues of covariances from standardised continuous quasi-identifiers of original and perturbed variables (eigenvalues can be estimated from a robust or classical version of the covariance matrix). This measure is mainly used to compare non-perturbative and perturbative methods. The greater the value of this measure, the

greater the changes in the dataset, and consequently, the greater the information loss (minimum value is 0).

A review of methods for the information loss assessment for categorical or continuous variables can be found in: Hundepool et al. (2010; 2012), Shlomo (2010), Templ, Kowarik and Meindl (2015), Templ (2017), Benschop, Machingauta and Welch (2019). The literature focuses on the problem of assessing information loss when perturbing categorical variables (methods are based on comparison of the variable's distribution before and after perturbation), and in situations when both types of variables are subjected to perturbation.

As regards information loss due to the use of perturbative methods, one of the objectives of the empirical study was to check how these methods affect point estimates and their quality. As mentioned in the introduction, the analysis was conducted for the unemployment rate, which is one of the key indicators based on LFS microdata (measured at the province level, total value and cross-classified by *Sex*). Estimates of this indicator together with estimates of the coefficient of variation were obtained using the PROC SURVEYMEANS procedure in the SAS programme (Lewis, 2016). In this procedure, the unemployment rate was determined as a weighted average of the binary variable, which took the value "1" for unemployed persons aged 15–74, and the value "0" for other economically active persons. More details on the Taylor series method – used in PROC SURVEYMEANS to estimate variance – can be found in Wolter (2007) and Lohr (2010).

5. The application of Statistical Disclosure Control methods in the Labour Force Survey

5.1. The confidentiality of original microdata

The empirical study was based on unit-level data from the LFS collected in the fourth quarter of 2011. Annual data were not used because of the structure of the survey sample: after combining quarterly unit-level datasets into an annual microdata set, some respondents may be included twice, which would result in underestimating the disclosure risk. The target unit-level dataset, limited to records with positive sampling weights, including persons aged 15 and older, contained 88,208 records.

In the microdata set, some variables were selected as categorical and continuous quasi-identifiers. The selection was based on their availability in external databases, which could potentially be available to the intruder. Seven categorical quasi-identifiers were identified: *Sex*, *Marital status*, *Labour market status*, *Level of disability*, *Level of education*, *Territorial unit code at NUTS4 level* and *Regis-*

tration at the employment agency as an unemployed person. The *Age* of the respondent was classified as a continuous quasi-identifier.

However, two of these categorical quasi-identifiers could not be perturbed. The *Labour market status* is determined on the basis of a few questions from the survey questionnaire. If one were to perturb only this one variable, the intruder could detect inconsistencies in the microdata and determine the real value of the variable based on the related questions. The *Territorial unit code* was also not perturbed because province codes are used to create identifiers for households, persons or strata. If values of the *Territorial unit code* were to be perturbed, one would have to exclude the possibility of replacing them with other values, belonging to another strata or province (because such inconsistency would attract the attention of the intruder and enable them to reproduce the true value of the perturbed variable).

It should be emphasised that in addition to the above-mentioned quasi-identifiers, available from many other sources, microdata made available for scientific purposes would include all variables collected in surveys – many of them obtained only in the LFS. This article focuses only on perturbing quasi-identifiers, but when preparing a unit-level dataset to be made available to users, it would be necessary to protect other variables, especially confidential ones.

After choosing quasi-identifiers, the confidentiality of the original microdata set was evaluated. The R programme report contains information about observations violating the *k*-anonymity principle, for *k* equals to 2, 3 or 5. For $k = 2$, the anonymity rule was violated by 17,361 observations (i.e. 19.68% of the sample), for $k = 3$ – by 27,665 observations (31.36%), and for $k = 5$ – by 40,851 records (46.31%). The result for $k = 3$ is important in the light of the *Public Statistics Act*, because, according to this law, only aggregates created from at least 3 observations can be published. This means that for 31.36% of observations, the same combinations of categorical quasi-identifiers do not occur at least three times in the dataset. The high percentage of observations violating this rule is related to the number of selected categorical quasi-identifiers and the large number of levels of the *Territorial unit code* (and small sample sizes in these cross-classifications).

The expected number of re-identifications is 391.40 (which accounts for 0.44% of the total sample size). The hierarchical structure of LFS microdata has to be considered when assessing the global risk and the expected number of re-identifications. The hierarchical structure – including personal and household identifiers – increases the risk of disclosure, because identification of even one household member will lead to the re-identification of other household members. When the household identifier is included in the risk assessment, the expected number of re-identifications increases to 1,060.42 (which represents 1.20% of the total sample size).

For values of the continuous quasi-identifier *Age*, the disclosure risk is between 0% and 100%.

5.2. The application of perturbative methods and their impact on the confidentiality and utility of microdata

Categorical quasi-identifiers were perturbed by applying the PRAM method while the continuous quasi-identifier – by means of Additive Noise or Rank Swapping. Thus, two perturbed datasets were obtained, one combining the results of PRAM and Additive Noise, and the other one, combining the results of PRAM and Rank Swapping.

In the empirical study, the invariant PRAM approach was used. The alpha parameter, responsible for the size of perturbations, was left at the default level of 0.5 (permissible values from 0 to 1).

In the LFS microdata set, non-responses occurred only in two categorical quasi-identifiers – in the *Level of disability* and the *Registration at the employment agency as an unemployed person*. Markov matrices were generated for each variable by the PRAM procedure. Table 1 presents the number and percentage of changes in each categorical variable.

Table 1. The number and percentage of records whose values for the categorical variables were perturbed by applying the PRAM method

Variable	Number of changes	Percentage of the sample
Sex	4,222	4.79
Marital status	4,400	4.99
Level of education	9,270	10.51
Level of disability	6,309	7.15
Registration at the employment agency	1,990	2.26

Source: own elaboration based on microdata from the LFS

After applying the inverse PRAM, the share of records violating the 3-anonymity increased by 3.70 percentage points. The percentage of expected re-identifications increased by 0.06 percentage points, and after taking into account the household identifier – by 0.20 percentage points. Because of using PRAM, some unique combinations of categorical quasi-identifiers in the perturbed microdata set may cease to be unique or may no longer exist, but completely new unique combinations may appear instead. For this reason, if at least one quasi-identifier is perturbed by PRAM, above measures of disclosure risk should not be interpreted. One can check whether the unique combination of categorical quasi-identifier values has changed for records with a high disclosure risk in the original microdata set. The R output does not contain any measure of information loss resulting from the perturbation of categorical variables. After perturbing categorical variables in microdata, one should always check whether the resulting combinations of values are logical or if certain patterns have been preserved (e.g. there are no observations

corresponding to a 15-year-old widow or a 15-year-old who declares holding higher education with a doctoral degree).

Despite the use of invariant PRAM, the subsequent limitation of the sample to persons aged 15–74 who are employed or unemployed (in total or by *Sex*) – for the purpose of estimation – disrupted the original sample structure.

In the empirical study, two perturbative methods for the *Age* variable were considered: Additive Noise (*method* = '*additive'*) with the amount of noise (in percent) – 10 (*noise* = 10); or by Rank Swapping with 1% grouping of the lowest and highest values of the variable before ranking (*TopPercent* = 1, *BottomPercent* = 1), with the multivariate preservation factor at 0.95 ($R_0 = 0.95$), with the subset-mean preservation factor (K_0) and the rank range (percentage of total sample size) (P) determined by the sample size, satisfying dependencies:

$$\frac{2K_0}{\sqrt{N}} = 5\%, \tag{3}$$

$$\frac{PN}{100} = 5, \tag{4}$$

where:
N – sample size,
$K_0 = 7.425$,
$P = 0.006$.

Regardless of the perturbative method used, the disclosure risk for the continuous variable *Age* does not decrease – it remains within the range of 0% to 100%. In the case of the original Rank Swapping method, values of the variable and their distribution do not change – only their order is changed. It is true in the case of censuses or administrative registers (which include all units); in the case of sample surveys this only holds for the sample – when the results are generalised to the population, the distribution may differ owing to differences in sampling weights. Similarly, in the case of Additive Noise – perturbing original values of the variable by adding random disturbances may result in the appearance of unique values (with high disclosure risks), especially when the number of possible values is unlimited. Perturbed values may also be insufficiently different from original ones.

In terms of information loss, in the case of Additive Noise, the measure *IL*1 is equal to 353,764.37, and in the case of Rank Swapping to 646,087.51. The other measure, Difference of Eigenvalues, is 0.00% for both methods. It can be concluded that a lower loss of information was observed when Additive Noise was used.

The following figures show the *Age* distribution of the target population in the 4th quarter of 2011, according to the original and perturbed microdata sets from the LFS, in total (Figure 1) and by *Sex* – for men (Figure 2) and for women (Figure 3).

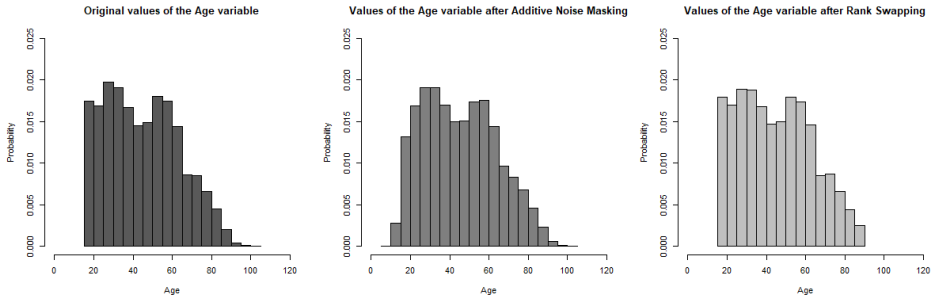

Figure 1. A comparison of the impact of perturbative methods on the *Age* distribution of the population aged 15+ in 4Q 2011

Source: own elaboration based on microdata from the LFS

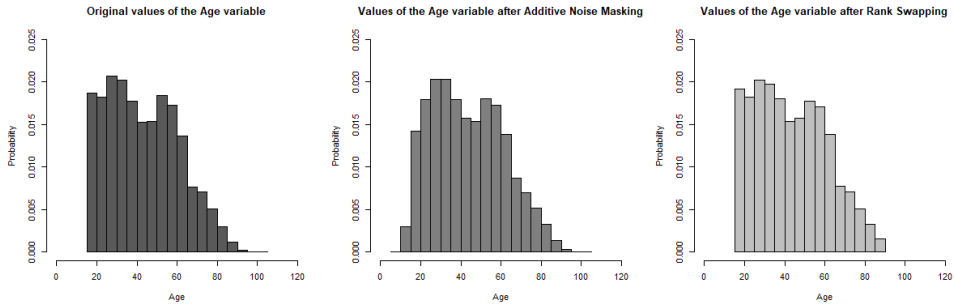

Figure 2. A comparison of the impact of perturbative methods on the *Age* distribution of the male population aged 15+ in 4Q 2011

Source: own elaboration based on microdata from the LFS

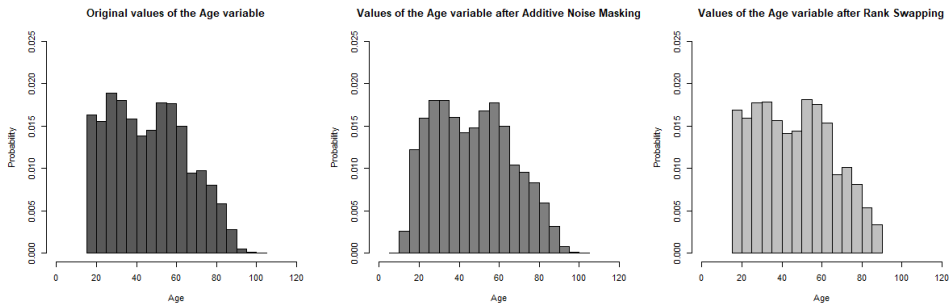

Figure 3. A comparison of the impact of perturbative methods on the *Age* distribution of the female population aged 15+ in 4Q 2011

Source: own elaboration based on microdata from the LFS

Age distributions obtained on the basis of perturbed datasets are similar to those that could be obtained from the original microdata. However, there are differences between them – bigger when the distribution is estimated separately for each sex, smaller for the total population. The biggest discrepancies can be observed in the tails of the distributions, which tend to include individuals who are at the risk of disclosure. The use of the Rank Swapping, in which a fixed percentage of observations with the lowest and highest values of the variable are grouped, leads to a reduced range of values. The Additive Noise, which introduces positive or negative disturbances, may increase the range of values.

As mentioned before, when using perturbative methods, the output microdata must be verified in order to check whether values of each perturbed variable are logical, and whether regularities between values of the variables have been preserved. Several inconsistencies were detected during the verification. Firstly, owing to the use of Additive Noise, values of the *Age* variable had a decimal part. Similarly, in the Rank Swapping – values in the tails of the distribution were replaced with their average value. For this reason, they had to be rounded off to the nearest integer. Secondly, after perturbing *Age* by Additive Noise, the value of the variable for 525 respondents was lower than 15. Such persons are not included in the LFS population, which is why it was set to 15 for all these respondents. Thirdly, according to the LFS definition, an unemployed person is defined as a person between the age of 15 and 74 fulfilling certain conditions. After perturbing the *Age* variable, the microdata set contained 2 persons (in the case of Additive Noise) or 1 person (in the case of Rank Swapping) with unemployed status and older than 74. In these case, the value of *Age* was changed to 74.

The above-presented inconsistencies are important from the estimation point of view. Other variables would have to be verified as well.

5.3. The estimation of the unemployment rate and its precision

The final stage of the empirical study was the estimation of the unemployment rate at the level of provinces, in total and by *Sex*, in order to indicate the impact of perturbations on point estimates and their quality. Figures 4 and 5 compare estimates (Figure 4) and values of the coefficient of variation for these estimates (Figure 5) obtained from the original and perturbed microdata.

Differences in the unemployment rate estimates at the province level (in total) were affected only by the *Age* perturbation. For this reason, estimates for the total population are not very different (of course, a relevant test of significance would have to be conducted to validate this conclusion). Bigger discrepancies occur when the unemployment rate is estimated for provinces cross-classified by *Sex*. The slight differences between estimates obtained from two perturbed datasets result from

the identical perturbation of categorical variables in both datasets. The analysis of estimation precision leads to the same conclusion – interestingly, the perturbation of values is not always associated with a fall in precision.

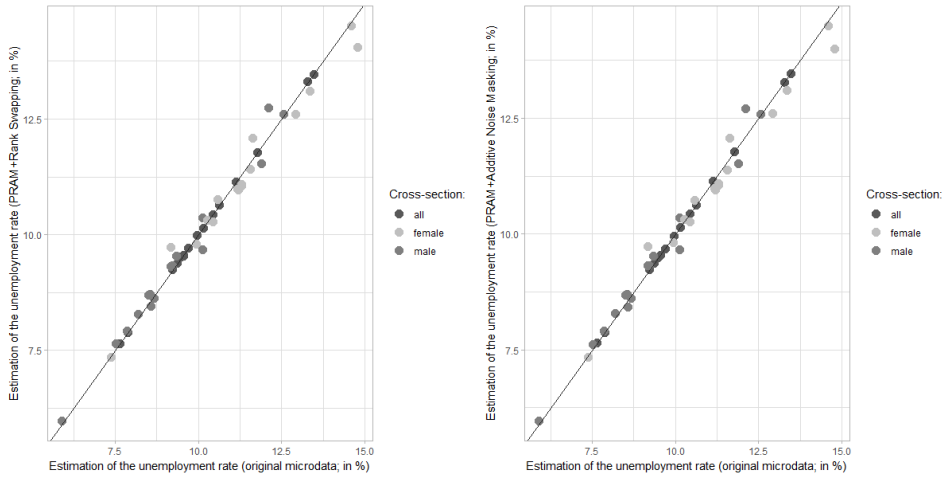

Figure 4. A comparison of the impact of perturbative methods on unemployment rate estimates in provinces, in total and by *Sex*, in 4Q 2011

Source: own elaboration based on microdata from the LFS

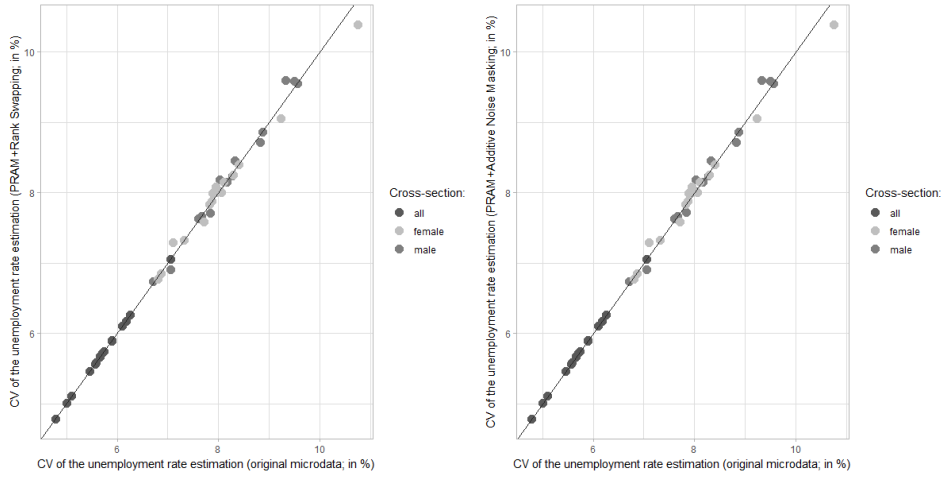

Figure 5. A comparison of the impact of perturbative methods on the coefficient of variation of unemployment rate estimates in provinces, in total and by *Sex*, in 4Q 2011

Source: own elaboration based on microdata from the LFS

The following figures show the distribution of estimates of the unemployment rate (Figure 6) and their coefficients of variation (Figure 7), in total and by *Sex*.

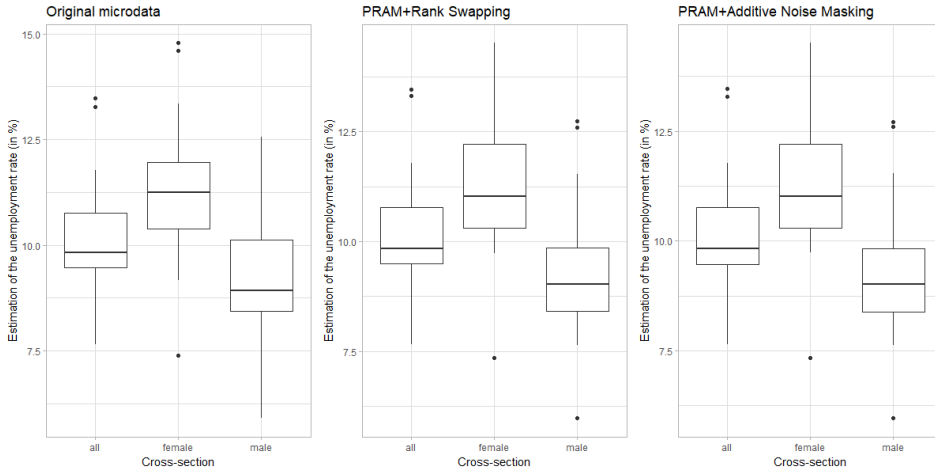

Figure 6. A comparison of the distribution of the unemployment rate estimates in provinces in 4Q 2011, in total and by *Sex*, depending on the microdata set used

Source: own elaboration based on microdata from the LFS

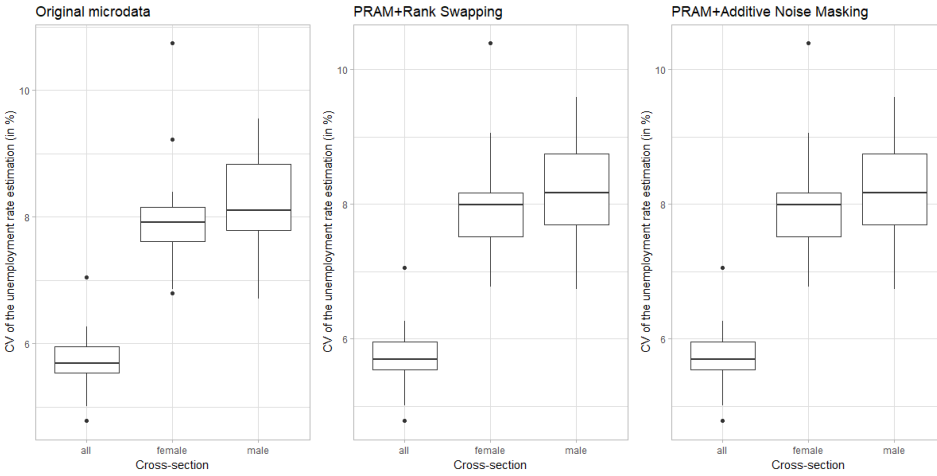

Figure 7. A comparison of the distribution of the coefficients of variation of unemployment rate estimates in provinces in 4Q 2011, in total and by *Sex*, depending on the microdata set used

Source: own elaboration based on microdata from the LFS

Distributions of the unemployment rate estimates and their coefficients of variation across provinces are similar. There are differences for each *Sex* category (e.g. in the mean, quartiles, minimum and maximum values).

6. Conclusions and further work

While the results of the first attempt to apply selected SDC methods to real microdata from the LFS are promising, a more thorough analysis of perturbed microdata is necessary – especially as regards their confidentiality and utility as well as the impact of applied methods on estimation quality; one should also examine inconsistencies in variable values or disturbed relations between variables. The accurate choice of methods and appropriate values of their parameters will have a crucial impact on the confidentiality and utility of microdata. It should be emphasised that LFS microdata are affected by sampling and non-sampling errors, and the application of SDC methods will be another source of error. Undoubtedly, a nonresponse is another important source of non-sampling error. The overall non-response rate for the LFS has recently exceeded 40%. In the literature, an attempt has been made to include the use of statistical disclosure limitation methods in the total survey error (Biemer et al., 2017). According to the author, it is necessary to develop an algorithm that makes it possible to detect quasi-identifiers, or at least propose some rules for their selection. The possibility of assessing the risk of disclosure and the loss of information in the *sdcMicro* package is limited. It also seems necessary to attempt a combined assessment of disclosure risk and information loss when detected quasi-identifiers are both categorical and continuous. Another requirement would be to set a maximum acceptable level of risk and the level of utility that microdata have to meet. Microdata for scientific purposes should be prepared once and made available to all interested persons in the same form. Each method presented in this article generates other disturbances in subsequent applications. A comparison of several versions of perturbed microdata could provide the intruder with some information about the size of perturbations. The appearance of new, unique combinations of categorical quasi-identifiers resulting from the use of perturbative methods may give the illusory impression of possible re-identification.

It seems that the application of SDC methods to microdata about the entire population (censuses, administrative registers, etc.) is more important than in the case of sample surveys in which not all persons are included in the sample (additional protection).

When developing the statistical disclosure control process of LFS microdata for Polish official statistics, the approach used by Eurostat can be treated as a point of reference. The EU-LFS is conducted in all EU Member States, 4 candidate countries and 3 countries of the European Free Trade Association. Eurostat provides microdata from this survey in the form of Scientific Use Files (datasets for all Member States, Iceland, Norway, Switzerland and the United Kingdom; available to units recognised as research entities after submitting the project proposal) and Public Use Files (datasets for selected EU countries; commonly available on the

Eurostat's website). More information on how Eurostat prepares microdata from the EU-LFS can be found in Eurostat (2019).

In further studies, the author wants to test these and other non-perturbative and perturbative methods in different settings. The author intends to explore the issue of parameter estimation based on protected microdata from the LFS. The impact of methods will be checked, e.g. on selected estimators of small area statistics (regression estimator, ratio estimator, Fay-Herriot model) and in the context of the model-randomisation approach. The author also intends to address the question of estimation quality. Another important aspect to be examined in future work is the impact of survey methodology on the SDC process (for example, the use of a rotational sample scheme).

References

Benschop T., Machingauta C., Welch M. (2019), *Statistical Disclosure Control: A Practice Guide*, https://readthedocs.org/projects/sdcpractice/downloads/pdf/latest/ (accessed: 13.03.2020).

Biemer P.P., Leeuw E. de, Eckman S., Edwards B., Kreuter F., Lyberg L.E., Tucker N.C., West B.T. (2017), *Total Survey Error in Practice*, "Wiley Series in Survey Methodology", Wiley, New Jersey.

CSO (2012), *Labour Force Survey in Poland. IV quarter 2011*, *Statistical Information and Elaborations*, Statistical Publishing Establishment, Warsaw, https://stat.gov.pl/cps/rde/xbcr/gus/pw_aktyw_ekonom_ludn_IVkw_2011.pdf (accessed: 13.03.2020).

Domingo-Ferrer J., Torra V. (2003), *On the connections between statistical disclosure control for microdata and some artificial intelligence tools*, "Information Sciences", no. 151, pp. 153–170.

Domingo-Ferrer J., Torra V. (2004), *Disclosure risk assessment in statistical data protection*, "Journal of Computational and Applied Mathematics", no. 164–165, pp. 285–293.

Duncan G.T., Elliot M., Salazar-González J.-J. (2011), *Statistical Confidentiality. Principles and Practice*, "Statistics for Social and Behavioral Sciences", Springer Science+Business Media, New York–Dordrecht–Heidelberg–London.

Eurostat (2019), *EU Labour Force Survey Database User Guide*, European Commission, https://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf (accessed: 13.03.2020).

Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Lenz R., Naylor J., Schulte Nordholt E., Seri G., Wolf P.-P. de (2010), *Handbook on Statistical Disclosure Control*, ESSNet SDC A Network of Excellence in the European Statistical System in the field of Statistical Disclosure Control, https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf (accessed: 13.03.2020).

Hundepool A., Domingo-Ferrer J., Franconi L., Giessing S., Schulte Nordholt E., Spicer K., Wolf P.-P. de (2012), *Statistical Disclosure Control*, "Wiley Series in Survey Methodology", Wiley, Chichester.

Lewis T.H. (2016), *Complex survey data analysis with SAS*, CRC Press, Taylor & Francis Group, Boca Raton.

Lohr S.L. (2010), *Sampling: Design and Analysis*, Second Edition, Brooks/Cole Cengage Learning, Boston.

Matthews G.J., Harel O. (2011), *Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy*, "Statistics Surveys", vol. 5, pp. 1–29, http://dx-.doi.org/10.1214/11-SS074

Shlomo N. (2010), *Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility*, "Journal of Privacy and Confidentiality", vol. 2(1), pp. 73–91, https://journalprivacy-confidentiality.org/index.php/jpc/article/view/584/567 (accessed: 13.03.2020).

Templ M. (2017), *Statistical Disclosure Control for Microdata. Methods and Applications in R*, Springer, http://dx.doi.org/10.1007/978-3-319-50272-4

Templ M., Kowarik A., Meindl B. (2015), *Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro*, "Journal of Statistical Software", vol. 67(4), pp. 1–36, http://dx.doi.org/10.18637/jss.v067.i04

Willenborg L., Waal T. de (2001), *Elements of Statistical Disclosure Control*, Springer Science+Business Media, New York.

Wolter K.M. (2007), *Introduction to Variance Estimation*, Second Edition, "Statistics for Social and Behavioral Sciences", Springer Science+Business Media, New York.

Metody kontroli ujawniania danych dla mikrodanych z Badania Aktywności Ekonomicznej Ludności

Streszczenie: Celem artykułu jest analiza możliwości wykorzystania wybranych zakłóceniowych metod kontroli ujawniania mikrodanych na przykładzie danych jednostkowych z Badania Aktywności Ekonomicznej Ludności. W pierwszym etapie oceniona została ochrona poufności informacji w oryginalnym zbiorze danych. Po zaaplikowaniu wybranych metod, zaimplementowanych w pakiecie *sdcMicro* programu R, przedmiotem dociekań stał się wpływ tych metod na ryzyko ujawnienia, poniesioną stratę informacji, a także na jakość estymacji określonych wielkości dla populacji. Podkreślone zostały pewne problematyczne aspekty praktycznego wykorzystania kontroli ujawniania danych, zaobserwowane podczas przeprowadzonej analizy.

Słowa kluczowe: kontrola ujawniania danych, metody zakłóceniowe, PRAM, addytywne dodawanie szumu, wymiana rang, mikrodane, Badanie Aktywności Ekonomicznej Ludności, pakiet *sdcMicro*

JEL: C18, H83, J20

Received: 2019-01-12; verified: 2020-03-02. Accepted: 2020-05-19