DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 4, 2015, pp. 737–761, http://dx.doi.org/10.1515/JOS-2015-0043

Quality Indicators for Statistical Disclosure Methods: A Case Study on the Structure of Earnings Survey

Matthias Templ[1]

Scientific- or public-use files are typically produced by applying anonymisation methods to the original data. Anonymised data should have both low disclosure risk and high data utility.

Data utility is often measured by comparing well-known estimates from original data and anonymised data, such as comparing their means, covariances or eigenvalues.

However, it is a fact that not every estimate can be preserved. Therefore the aim is to preserve the most important estimates, that is, instead of calculating generally defined utility measures, evaluation on context/data dependent indicators is proposed.

In this article we define such indicators and utility measures for the Structure of Earnings Survey (SES) microdata and proper guidelines for selecting indicators and models, and for evaluating the resulting estimates are given. For this purpose, hundreds of publications in journals and from national statistical agencies were reviewed to gain insight into how the SES data are used for research and which indicators are relevant for policy making.

Besides the mathematical description of the indicators and a brief description of the most common models applied to SES, four different anonymisation procedures are applied and the resulting indicators and models are compared to those obtained from the unmodified data. The disclosure risk is reported and the data utility is evaluated for each of the anonymised data sets based on the most important indicators and a model which is often used in practice.

Key words: Statistical disclosure control; data utility; quality indicators; R.

1. Introduction

Anonymisation methods are applied to microdata to reduce their disclosure risk. By applying too much or overly heavy anonymisation, the data utility is reduced and the information loss is increased. However, users who analyse anonymised microdata want to have as precise parameter estimates as possible. It is therefore of great interest to measure the data and user context utility of a microdata set after disclosure limitation methods have been applied.

1.1. General Methods for Measuring Data Utility

Anonymised data should have the same structure as the original data and should allow for analysis with high precision.

[1] Statistics Austria, Dept. of Methodology, Guglgasse 13, Vienna 1110, Austria. Email: matthias.templ@gmail.com

Acknowledgment: I wish to thank three anonymous referees for their constructive reviews. My special thanks goes to the corresponding associate editor who improved the quality of the article.

To evaluate the precision, the estimation of different classical estimates such as means and covariances are often focused upon. By using the R-package **sdcMicro** (Templ and Meindl 2010; Templ 2008; Templ et al. 2015), it is possible to calculate 26 different measures on continuous scaled variables that are based on classical (most of these measures are described in Hundepool et al. 2012) or robust distances. These measures are computed for the original data and the perturbed data and then compared. To evaluate the multivariate structure of perturbed data, comparisons based on eigenvalues and robust eigenvalues may also be made. The comparison of means and covariances by mean squared errors, mean absolute errors, and mean variations is also proposed in Domingo-Ferrer et al. (2001). A generalisation is given by Domingo-Ferrer and Torra (2001) by averaging the mean variations and mean absolute errors. They also define information loss measures for categorical variables: direct comparison of categorical values, comparison of contingency tables, and entropy-based measures. For the direct comparison, a distance is defined over the range of categories. When the range of categories of a variable is of ordinal scale, the distance between two categories is proportional to the number of categories between them. For nominal scale, the Hamming distance (zero when equal, otherwise one) is chosen. The comparison of contingency tables considers the number of differences between the two contingency tables, normalised by dividing by the number of cells of a table. The entropy-based measure is suitable for the PRAM method, where the logarithm of the transition probabilities of one category to another is used.

Shlomo (2008) uses some methods to evaluate data utility based on a contribution from Gomatam and Karr (2003) and extends them by measures on impact of association and a measure based on the between variance of a proportion fitted by regression models.

Woo (2009) proposes the use of propensity-score methods. The idea is to merge or join the original and the perturbed data sets and then create a new index variable with ones for the original data and zeros for observations from anonymised data. A logistic regression model is then fitted using the new index variable as the response variable. Predictions from this model are then compared with the proportion of observations of the perturbed data to the original data (usually 1/2). Woo also describes two other measures, one based on cluster analysis (evaluating the cluster sizes) and another which compares the empirical cumulative distribution function. They concentrate only on data utility measures and do not account for disclosure risk. Karr et al. (2006) propose measures based on differences between inferences on original and perturbed data that are tailored to normally distributed data, and they also use the propensity score method in Oganian and Karr (2006).

Reiter (2012) mentions, without presenting numerical results, that the comparison of measures based on specific models is often done informally. If the regression coefficients obtained from original and perturbed data are considered close, for example if the confidence intervals obtained from the models largely overlap, the released data have high utility for that particular analysis (see also Karr et al. 2006).

1.2. Trade-Off Between Data Utility and Disclosure Risk

The goal of statistical disclosure control is always to release a safe microdata set with high data utility and a low risk of linking confidential information to individual respondents.

Disclosure risk can be measured in different ways. Several methods have been suggested, such as the individual risk approach (Franconi and Polettini 2004) that is used in this contribution, methods based on log-linear models (Rinott 1990; Carlson 2002) or the SUDA concept (Manning et al. 2008). So, firstly, a decision on which method, or methods, for measuring disclosure risk will be used is necessary. Secondly, the data holder has to decide on the level of disclosure risk that is acceptable and sufficient for distributing the data. For example, in the case of the SES, anonymised microdata is sent to Eurostat. However, many countries do not agree with the proposed rules for anonymisation communicated by Eurostat, nor can they allow the use of remote access systems such as the PiEP Lissy project (Marsden 2010) because of restrictions in national legislation. Therefore, almost every country applies different anonymisation methods to their data (the anonymisations and the disclosure risk are therefore somehow fixed in advance), but Eurostat wants to ensure that the most important statistics can be estimated with high precision.

In this study, the focus is not specifically on disclosure risk, however, and hence only one disclosure risk measure, the individual risk approach, was used. Several anonymisation procedures were however applied to the data and the data utility for each case is reported. It is up to the data holders to decide whether a particular anonymisation procedure is sufficient. In this study we have simply assumed that the chosen anonymisations are sufficient from a risk point of view and devote our attention to data utility.

1.3. Outline of the Article

In Section 2 we describe the basic ideas of the proposed approach for utility assessment. Section 3 introduces the Structural Earnings Survey (SES). In addition, the usage of this particular survey is analysed and the most important projects which have their main focus on this data set are mentioned. Based on this analysis, the most important indicators are discussed in Section 4 and the three most important indicators and one model are described in detail in Section 5. Confidentiality aspects are briefly discussed in Section 6. Results from the analysis using the selected data utility measures are presented in Section 7. Section 8 concludes the article.

2. Data and Context-Driven Utility Measures

In practice it is not possible to create an anonymised file that has exactly the same structure as the original file. Contrary to general methods described previously, we propose that the differences between estimates based on anonymised and original data need to be small, or even zero, only for the most important statistics. This approach measures the data utility based on quality indicators (Ichim and Franconi 2010; Franconi et al. 2011; Templ 2011a) and is another more user-driven approach than applying general tools, since for the users it might not be relevant to estimate all popular statistics with high precision, but just those that are relevant for their analysis.

The first step in quality assessment is to decide on a set of quality indicators. To do so, one has to evaluate the user needs, that is, what is analysed by the users, and report on the most important estimates. These estimators are often named *benchmarking indicators* (see, e.g., Templ 2011a,b) and referred to here as quality indicators.

The general procedure is quite simple – although much work is necessary. It can be described in the following steps:

i) Analysis of the user needs of researchers, policy makers, and society regarding a specific data set. Analysis of the aim for which the underlying data have been used.
ii) Selection of a set of quality indicators after the detailed analysis in (i).
iii) Estimation of all quality indicators on the original, unmodified microdata set.
iv) Estimation of the quality indicators on the protected microdata set.
v) Comparison of statistical properties such as point estimates, variances or overlaps in confidence intervals for each quality indicator.
vi) Assessment of the data utility of the protected microdata set.

If the quality of the data is reasonable, the anonymised microdata set may be published. Note that the anonymisation procedure chosen has to lead to a reasonably low disclosure risk of the anonymised data.

If the deviations of the main indicators calculated from the original and the protected data are too large, the anonymisation procedure should be revised by modifying selected parameters used for the applied disclosure methods or by a complete revision of the anonymisation process.

Usually the evaluation is focused on the properties of numeric variables given unmodified and modified microdata. However, it is of course also possible to look at the impact of local suppression or recoding that has been conducted to reduce individual reidentification risks.

Another possibility to evaluate the data utility is to define a model that is fitted to both, the original, unmodified microdata and the anonymised data. The main idea is to look at differences in the regression coefficients. If the deviations are small enough, one may go on to publish the safe and protected microdata set. Otherwise adjustments in the protection procedure need to be carried out.

It may also be of interest to evaluate the set of quality indicators not only for the entire data set but also for some domains. The evaluation of quality indicators is then performed for each of the h groups by looking at differences between indicators for original and modified data in each group.

3. The Structural Earnings Statistics Survey

The Structural Earnings Statistics Survey (SES) is conducted in almost all European countries, and the most important figures are reported to Eurostat.

3.1. Sampling Design, Data Preparation Issues, and Data Sources

SES is a complex survey of enterprises and establishments with more than ten employees (e.g., 11,600 enterprises in Austria), NACE C-O, including a large sample of employees (e.g., in Austria: 199,909). In many countries, a two-stage design is used where in the first stage a stratified sample of enterprises and establishments on the NACE one-digit level, NUTS 1 and employment size range is used – large companies have higher inclusion probabilities. In stage two, systematic sampling is applied within each enterprise using unequal inclusion probabilities with regard to employment size-range categories.

In the Austrian case, for example, the sample has only 2.4% nonresponse. Regression imputation is applied by using tax data to replace these missing values. If information on imputed values is available, variance estimation procedures should account for this extra variability.

Calibration is applied to reflect certain population characteristics corresponding to NUTS 2 and NACE one-digit level, but also for gender (number of men and women in the population).

SES compromises information from different perspectives and sources:

Information on the enterprise level: Enterprises are asked question batteries, such as whether the enterprise is private or public or whether it has a collective bargaining agreement (both binary variables). As a multinomial variable, the type of collective agreement is included in the questionnaire.

Information on the individual employment level: The following questions to employees come with the standard questionnaire: social identity number, date of employment, weekly work hours, kind of work agreement, occupation, amount of annual leave, place of work, gross earnings, earnings for overtime, and amount of overtime.

Information from registers: All other information may come from registers, such as information about age, size of enterprise, occupation, education, amount of employees, NACE, and NUTS classifications.

3.2. Standard Publications and Use of the Microdata

The standard publication from national statistical offices is issued every four years after the survey is conducted. In addition, a special publication about low incomes and non-common occupation employment is published by some member states, such as Statistics Austria's report on low incomes (see Geissberger and Knittler 2010). In Austria, a special report has been written for the Austrian women's report focused on the gender pay gap and socioeconomic studies (Geissberger 2010). Many other national publications by statistical agencies or researchers are available in almost every country (for some summaries about publications until 1999, see Belfield 1999; Nolan and Russel 2001; Dupray et al. 1999; Frick and Winkelmann 1999; Dell'Aringa et al. 2000).

However, social scientists have mostly carried out qualitative analysis or rough quantitative interpretations of a few official figures, mainly because of lack of access to micro data for researchers. One exception are publications made with direct or follow-up data connection and using the PiEP Lissy project and its remote access system (Marsden 2010) to various SES data. Actually, 10-15 projects are running within Eurostat's Safe Center and anonymised CD-ROM (see the next section).

3.3. Access to SES Microdata and European Projects

Access to Data Provided by Eurostat: Anonymised SES 2002 and 2006 data from 23 countries can be accessed for research purposes by means of research contracts through the safe center or anonymised CD-ROM at the premises of Eurostat. The output will be checked by Eurostat for confidentiality and quality. Further plans include automatic

output checking of data to reduce the workload of the statistical institutes. More technical details on the safe center can be found in Reuter (2010); and Reuter and Museux (2011). To obtain the data, see Eurostat's website: http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/ses.

Access to Data Through PiEP Lissy: The *Pay Inequalities and Economic Performance Project* (PiEP) studied wage differentials based on SES data (Marsden 2010) in depth. SES microdata from the Czech Republic, Hungary, Ireland, Italy, Latvia, Lithuania, the Netherlands, Norway, Portugal, Slovakia, and Spain can also be analysed via the PiEP Lissy remote-access system. The user can run Stata code on the PiEP Lissy server, for example, although some commands (twelve in total) are blocked by the system to prevent listing of individuals.

Synthetic SES Population Data: A synthetic population is simulated in Templ and Filzmoser (2014) and a sample of this population is included in the R-package laeken (Alfons and Templ 2013).

The LEED Project: Within the EU project on *Linked Employer-Employee Data* (LEED), studies assessing the potential of linked employer-employee and panel data sets for the analysis of European labour-market policy are carried out. They concentrate on SES data and use the PiEP Lissy remote access system to gain access to the data of twelve different countries, see http://cep.lse.ac.uk/leed/.

The Dynamic Wage Network: The dynamic wage network was founded by the European Central Bank and it consists of four research groups. The microdata group pursues three directions of research one of which is on wage differentials and modelling of earnings. The SES data is one of the main data sources for this group, used by many authors (see, e.g., Caju et al. 2010, 2009a,b; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010).

4. Important Indicators Estimated from SES Data

4.1. Research Potential of SES Microdata

Statistical agencies usually provides, amongst other things, tables on average hourly earnings on domain level (Geissberger 2009), for country comparisons (see, e.g., Mittag 2005) and for special groups like low incomes (Geissberger and Knittler 2010; Casali and Alvarez 2010).

SES data includes information on enterprise and employment level. Generally such linked employer-employee data are used to identify determinants/differentials of earnings, some indicators are also directly derived from hourly earnings, such as the gender pay gap or the Gini index (Gini 1912). The most classical example is the income inequality between genders as discussed in for example, Groshen (1991).

A correct identification of factors influencing earnings could lead to relevant evidence-based policy decisions. Research studies are usually focused on examining the determinants of disparities in earnings. Earnings comparisons between different industries or regions are frequently performed (see, e.g., Stephan and Gerlach 2005; Caju et al. 2010, 2009b,a; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010). Sometimes socioeducational factors are investigated as possible

explanatory variables of income, for example in Bowles et al. (2001). The overview of the analyses performed using SES data highlighted that, generally, the log hourly earnings are modelled. The explanatory variables correspond to employer activity (related to the enterprise), his or her experience (education, length of stay in service, qualification, etc.) and working hours. It was also observed that linear models are extensively used. ANOVA analysis, linear mixed-effects models, and multi-level models are other examples of statistical tools that have been applied. However, a lot of similar models are applied in the literature to model the log hourly earnings.

It should also be noted that the distribution of errors is always assumed to be normal. The estimates are generally computed by means of ordinary least squares by ignoring the sampling design and corresponding weights which is not good practice.

4.2. Summary of the Most Important Analyses from SES Data

In summary, the most important analyses using SES data are related to

Gender pay/wage gap: The gender wage gap is currently one of the most important indicators obtained from SES in many European countries (Research Center for Education and the Labour Market at the Maastricht University 2009) and intensively discussed in the European Union (Dupré 2010). In Austria, for example, many publications about the gender wage gap are published by Statistics Austria and the national authorities (Stockinger 2010). The topic *Women and Equality* is of central interest not only for the Federal Minister for Women and the Civil Service, and socioeconomic studies are carried out with support from the state (one example is Geissberger 2010) or European institutions where regression models are also applied to estimate the adjusted gender pay gap (Research Center for Education and the Labour Market at the Maastricht University 2009).

Wage differentials and interindustry wage differentials: Differences in earnings for workers employed in different industries and occupations has long been recognised as an important issue for the labour market and several studies have been carried out (Caju et al. 2010, 2009a,b; Messina et al. 2010; Dybczak and Galuscak 2010; Simón 2010; Pointner and Stiglbauer 2010). Pointner and Stiglbauer (2010) use several workplace-specific dummy variables for the employee's occupation (ISCO 1) within the firm, the sector (NACE-2 digits) of the employer, for firm size and location (NUTS-1 digits), and a control for private ownership of the firm as predictors. Caju et al. (2010, 2009b) modelled the gross hourly wages with sex, education, age class, number of years of employment, type of employment contract, part/full-time, bonus for shift work, night and/or weekend work, a dummy for paid overtime and occupation sector effect. Messina et al. (2010) used a model to predict the log hourly wages with firm size, firm size squared, age class, female employment proportion and proportion of high- and low-skilled workers as predictors. Caju et al. (2009a) used age, capital-labour ratio, profit elasticity and the percentage of blue-collar workers covered by single-employer collective agreements to model the log hourly earnings.

Low-pay dynamics: In some countries, great changes in the distribution of earnings are observed (see, e.g., Dell'Aringa et al. 2000; Geissberger 2009) with a widening of

inequality and an increase in dispersion. The Gini index and the quintile share ratio are two of the main indicators to estimate the inequality (Graf et al. 2011; Kolb et al. 2011).

Enterprise characteristics that affect earnings or profit: The differential that describes the profit of an enterprise is an interesting aspect, that is how enterprises integrate a combination of systems to provide greater flexibility in pay, and how information sharing and the size of the enterprise influences the profitability of an enterprise. On the other hand, it is of interest to investigate the prediction of pay flexibility using the size of the enterprise, level of competition, training, job rotation, time flexibility, and so on (see, e.g., Marsden 2010).

Collective bargaining: Due to the unions importance in determining wages, to measure the extent of the union-nonunion wage gap is of interest (for an example from the US, see Edwards 2010; also see Fitzenberger et al. 2006).

Average Earnings: Average earnings in enterprises as an indicator for productivity or performance (Winter-Ebmer and Zweimüller 1999; Marsden 2010). The idea is that in a competitive market environment, employees' pay corresponds to the value of their output, that is deviations from this position would lead to difficulties in recruitment and retention. In branches with high output, earnings would therefore be higher compared to enterprises in low economic branches with low production.

Occupation and length of employment: Another interesting analysis includes the difference in income for different occupation levels or by the length of employment.

Comparative studies between countries play an increasingly important role. However, our purpose is to study how estimates of a defined set of indicators from protected microdata perform compared to estimates based on the original, unmodified data. Therefore, such comparative studies are not directly within the scope of this work, since good estimates on a country level should ensure that comparisons between countries are possible.

5. Two Indicators and One Model for Quality

In the following, three measures that we have identified as the most important and have selected as quality indicators are described in full detail. Note that in a real-life setting, one would include any number of measures deemed important enough and not just the three we have chosen. However, in order to avoid this article becoming overly long, we limit the investigation to only these three quality statistics.

First, the (unadjusted) gender pay gap (GPG) is described, since it is one of the most important indicators obtained from SES data; thereafter the Gini index is described. The GPG and the Gini index (for hourly earnings) are extremely sensitive to changes in the upper and lower tail of the distribution (see e.g., Alfons et al. 2013). If these estimators are not affected by anonymisation, one can be quite sure that the corresponding variables have high data utility, since it is most difficult to preserve the structure of the data in the upper tail of the distribution.

Lastly, a model-based estimation on employment level is described, representative for all model-based estimations. Note that our choice of indicators and model is subjective; even so, the choice is based on our review of dozens of contributions (see Subsection 4.1). However, it can be expected that differences in estimations between anonymised and original data according to this model will be comparable in similar models.

5.1. The Gender Pay Gap

As already noted, the GPG is probably the most important indicator derived from the SES data.

The calculation of the GPG is based on each person's hourly earnings. The hourly earnings equals to the gross monthly earnings from labour divided by the number of hours usually worked per week during 4,33 weeks, (see EU-SILC 2009; Beblot et al. 2003).

5.1.1. Definition Gender Pay Gap

The GPG in unadjusted form is defined on population level as the difference between average gross earnings of male paid employees and of female paid employees divided by the earnings of male paid employees (EU-SILC 2009).

5.1.2. Estimation of the Gender Pay Gap

Since the GPG is usually estimated by survey information, the estimation has to consider sampling weights in order to ensure sample representativity. Therefore, all our estimations consider sampling weights.

We let $\boldsymbol{x} := (x_1, \ldots, x_n)'$ denote the hourly earnings where $x_1 \leq \ldots \leq x_n$ and $\boldsymbol{w} := (w_i, \ldots, w_n)'$ denote the corresponding personal sample weights, where n denotes the number of observations.

We define the index set

$$J^{(M)} := \{j \in \{1, \ldots, n\} \mid \text{worked as least 1 hour per week} \wedge (16 \leq \text{age} \leq 65) \wedge \text{person is male}\},$$

and let $J^{(F)}$ be the corresponding index set for female employees.

With these index sets, the GPG in its unadjusted form is estimated by

$$GPG_{(mean)} = \frac{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i} - \frac{\sum_{i \in J^{(F)}} w_i x_i}{\sum_{i \in J^{(F)}} w_i}}{\frac{\sum_{i \in J^{(M)}} w_i x_i}{\sum_{i \in J^{(M)}} w_i}}. \quad (1)$$

The definition from EU-SILC (2009) differs from the definition used by the Bureau of Labour Statistics of the United States (see, e.g., Weinberg 2007), where weighted medians are used instead of arithmetic means.

The GPG is usually estimated at domain level such as economic branch, education and age groups (Geissberger 2009).

In addition, it is important to estimate the variances of the estimations.

5.2. The Gini Index for the Estimation of Inequality

The Gini index (Gini 1912) is a well-known measures of inequality of a distribution and is widely applied in many fields of research.

The Gini index according to EU-SILC (2004, 2009) is estimated by

$$\widehat{Gini} := 100\left[\frac{2\sum_{i=1}^{n}\left(w_i x_i \sum_{j=1}^{i} w_j\right) - \sum_{i=1}^{n} w_i^2 x_i}{\left(\sum_{i=1}^{n} w_i\right)\sum_{i=1}^{n}(w_i x_i)} - 1\right]. \tag{2}$$

The Gini index is closely related to the Lorenz curve (Lorenz 1905), which plots the cumulative proportion of the total income against the corresponding proportion of the population.

The Gini index and the GPG are typically – among other domains – estimated with breakdowns by age and gender, or age, gender, and region, or by education level. The latter domain is used in the following.

5.3. *Model-Based Predictions on Employment Level*

As representative of all model-based estimations at employment level, we choose a model described in Marsden (2010) applied within the PiEP Lissy project and also used in Dybczak and Galuscak (2010). They fit OLS regression models where they modelled the gross hourly earnings of workers in enterprises using age, age^2, sex, education, and occupation as predictors.

The data from the Lissy system is also used for the LEED project (see Subsection 3.3) where similar studies and modelling have been carried out (see, e.g., Simón 2010). Similar models are also fitted within the *wage dynamics network* of the European Central Bank (Caju et al. 2010; Pointner and Stiglbauer 2010).

In the following estimations, the following model is used:

$$\begin{aligned} log(\text{hourly earnings}) \sim\ & \text{sex}\,(2) + \text{age}\,(6) + \text{education}\,(6) + \text{occupation}\,(23) \\ & + \text{location}\,(5) + \text{economic activity}\,(12) + \text{error term} \end{aligned}$$

The numbers in brackets correspond to the respective number of categories for each of the categorical variables in the original SES data.

It seems that the sampling weights are mostly ignored in the literature on fitting models to SES data. However, in our study the weights are taken into account by using weighted least squares regression.

5.4. *Variance Estimation*

A calibrated bootstrap to estimate the variances (Bruch et al. 2011; Templ and Alfons 2011) for the GPG and the Gini index is applied.

Let X denote a survey sample with n observations and p variables. Then the *calibrated bootstrap algorithm* for estimating the variance and confidence interval of an indicator can be summarised as follows:

1. Draw R independent bootstrap samples $\boldsymbol{X}_1^*, \ldots, \boldsymbol{X}_R^*$ from $\boldsymbol{X}$.

2. Calibrate the sample weights for each bootstrap sample $\boldsymbol{X}_r^*$, $r = 1, \ldots, R$. Generalised procedures are then used for calibration: a multiplicative method known as *raking*, an additive method or a logit method (see Deville and Särndal 1992; Deville et al. 1993).
3. Compute the bootstrap replicate estimates $\hat{\theta}_r^* := \hat{\theta}(X_r^*)$ for each bootstrap sample X_r^*, $r = 1, \ldots, R$, where $\hat{\theta}$ denotes an estimator for a certain indicator of interest. The sample weights need to be considered in the computation of the bootstrap replicate estimates.
4. Estimate the variance $V(\hat{\theta})$ by the variance of the R bootstrap replicate estimates:

$$\hat{V}(\hat{\theta}) := \frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{\theta}_r^* - \frac{1}{R} \sum_{s=1}^{R} \hat{\theta}_s^* \right)^2 \tag{3}$$

5. Estimate the confidence interval at confidence level $1 - \alpha$ by the percentile method: $\left[\hat{\theta}^*_{((R+1)\frac{\alpha}{2})}, \hat{\theta}^*_{((R+1)(1-\frac{\alpha}{2}))} \right]$, as suggested by Efron and Tibshirani (1993), where $\hat{\theta}^*_{(1)} \leq \ldots \leq \hat{\theta}^*_{(R)}$ denote the order statistics of the bootstrap replicate estimates.

6. Confidentiality Issues and Perturbation of SES

6.1. Disclosure Scenario

In principle, two reidentification scenarios are related to the SES data. The identification of an enterprise may lead to information about their employees. Key variables at enterprise level might be *location* (3), NACE one-digit level codes (*economic activity*) (12), *size* of the enterprise (5), and distinction between *public or private* enterprises (2); the bracketed numbers are the respective number of categories. However, here we only focus on reidentification scenarios on employment level since the fraction of employees asked in each company, is rather high (lower for large enterprises, larger to all employees in smaller companies). Furthermore, to limit the scope of the paper, more serious disclosure situations on employment level will not be considered.

Categorical key variables at employment level might be *location* (3), *age class* (6), *education* (7), *economic activity* (12), and *size* (5). This leads to 7,560 strata. Of course, the choice of key variables for disclosure scenarios is a somewhat subjective decision and might vary across countries. For example, Ichim and Franconi (2007) proposed to use only *location*, *economic activity*, *size* and *age class* as categorical key variables. Continuous key variables at employment level might be the *hourly earnings* and *overtime earnings*. This choice of scenario is also a subjective decision.

Remark: Anonymised SES 2002 and 2006 data from 23 countries can be accessed for research purposes through the safe center at the premises of Eurostat. Anonymisation is done by recoding NACE, NUTS, and size, removing citizenship and building six age classes, microaggregation (individual ranking) for absence days and earnings and removing the sampling weights.

6.2. *Anonymisation of SES*

Various methods exists to anonymise microdata (see, e.g., Hundepool et al. 2012; Templ and Meindl 2010). Two possibilities (amongst others) for anonymisation are the following:

a) To provide *k*-anonymity (Sweeney 2002) for categorical key variables (for enterprises, for employees), and to apply microaggregation or adding (correlated) noise (Brand 2004) for continuous key variables.
b) Synthetic data generation of all variables (Alfons et al. 2011; Templ and Filzmoser 2014), that is, simulation of all variables by drawing from predictive distributions. Note that by simulating only a part of variables (e.g., gross earnings) and leaving other variables (such as the categorical variables) unchanged, intruders might be able to identify persons based on the unchanged variables and this might not be in scope with specific legislations on data privacy.

Fixed rules to protect the microdata may not always be accepted by all data providers (e.g., member states of EU); some freedom to choose protection methods must be given. However, some minimal quality requirements must be fulfilled by the applied protection methods (Ichim and Franconi 2010).

We do not go into detail about the anonymisation methods *per se* since the main focus of this paper is on evaluating the data utility of anonymised data.

Nevertheless, three possible perturbations to make the data confidential are outlined and applied. First, variables *size*, *age*, *sex*, *location*, *education* and *economic activity* are selected as categorical key variables and *hourly earnings* and *overtime earnings* as continuous key variables. Then the following anonymisation procedures are applied (note that this choice of anonymisation methods is subjective and many other disclosure scenarios and perturbation methods can be applied):

1. Recoding from 53 categories to twelve categories for the variable *economic activity*: local suppression to achieve three-anonymity (optimal local suppression following Templ et al. 2015); microaggregation (individual ranking method for fast computations) applied on each strata defined by *economic activity* of *hourly earnings* and *overtime earnings* with aggregation level 4.
2. Same recoding and local suppression as in 1: adding correlated noise (Brand 2004) to *hourly earnings* and *overtime earnings* with noise parameter 150 (for details, see Templ et al. 2015).
3. Swapping *location* and *economic activity* using the (invariant) postrandomization method (PRAM, see Gouweleeuw et al. 1998) with default parameters (see Templ et al. 2013); microaggregation as in 1.
4. Experimentally, shuffling (Muralidhar and Sarathy 2006) with a rather small model is applied (earnings hour + earnings overtime $\sim$ sex + age + education); the anonymisation of categorical key variables are done as in 1 (and 2).

The amount of local suppression (to achieve three-anonymity) for Procedures 1, 2 and 4 is 0.001% (one value out of 199,909) for *size*, 0.115% (230 values) for *economic activity* and 0.005% (nine values) for *age*.

Table 1. Comparison of different anonymisation methods using the absolute relative bias (arb) for the GPG and the Gini indices. Overall indicates the estimation of bias without taking domains into account; for GPG, arb over domains education and age is calculated; for the Gini index, arb over domain age × sex is calculated. Global disclosure risk and the number of expected reidentifications are reported in the last two columns, specifically the amount of observations violating three-anonymity (three-anon), the sum of individual risks (grisk) and the expected numbers of observations disclosed (for the latter two, see Franconi and Polettini 2004). Asterisks (for PRAM) indicate that the estimates might not be reasonable

	Gender pay gap			Gini index		Global risk		
	Overall	Education	Age	Overall	Age-sex	3-anon	Grisk	Ident.
original	–	–	–	–	–	4414	0.010	2024
rec + ls + ma	0.176	0.671	0.861	0.081	0.191	0	0.002	426
rec + ls + noise	0.669	0.524	2.478	0.646	1.484	0	0.002	426
pram + ma	0.010	0.185	0.314	0.001	0.100	1011*	0.003*	539*
rec + ls + shuffle	11.918	18.677	152.381	0.009	18.856	0	0.002	426

For the application of PRAM in Procedure 3, 18,151 values changed their category in *location* and 18,867 values in *economic activity*.

7. Results

The utility measures chosen – based on the quality indicators that have been defined in Section 5 – are the following:

- The difference in the estimation of the GPG and the Gini from the original and perturbed data defined for h domains given by the (well-known) absolute relative bias:

$$arb = \frac{1}{h}\sum_{i=1}^{h}\frac{\left|\hat{\theta}_i - \tilde{\theta}_i\right|}{\hat{\theta}_i}, \tag{4}$$

- where $\hat{\theta}$ and $\tilde{\theta}$ denote the estimates from the original and the anonymised data set respectively. Note that the $\hat{\theta}$ have to be nonzero, which is practically always the case.
- The variances are estimated and the overlap of the confidence interval of the perturbed and original data is evaluated and reported as percentages.
- The model defined in Subsection 5.3 is fitted using weighted least squares regression on original and perturbed data. To stay comparable, the categories of *economic activitiy* are equal, that is, the NACE one-digit level is chosen.

7.1. Absolute Relative Bias

Table 1 shows the absolute relative bias (arb) for the GPG and the Gini index; both the overall estimate and the mean over the domains is shown. Here, the domain *education* and *age* is chosen for the GPG and for the Gini index, the domain (sex × age class) is used since this is reported to be one of the most interesting domains (see, e.g., Geissberger 2009, EU-SILC 2009 and Section 5).

The global measure of individual risk and the expected number of reidentifications are reported in the last two columns of Table 1. Note that the sum over individual risks gives the number of expected reidentifications. The number of reidentifications is not high in the original data set (2,024 of 199,909 observations) and it is reduced by applying the

Table 2. Lower (l) and upper (u) limits of the confidence intervals for the GPG for each category of education

Data	ISCED 0–1	ISCED 2	ISCED 3–4	ISCED 5A	ISCED 5B
original (l)	0.15938	0.12102	0.22572	0.29568	0.21744
original (u)	0.26525	0.15023	0.23944	0.35010	0.25835
rec + ls + ma (l)	0.16123	0.12144	0.22624	0.28891	0.21290
rec + ls + ma (u)	0.27062	0.15211	0.23970	0.34381	0.25904
rec + ls + noise (l)	0.17012	0.12106	0.22399	0.29135	0.21152
rec + ls + noise (u)	0.27011	0.15172	0.23776	0.34551	0.25805
pram + ma (l)	0.17682	0.12200	0.22554	0.29064	0.21946
pram + ma (u)	0.27230	0.15065	0.24197	0.33822	0.26172
rec + ls + shuffle (l)	−0.01865	0.09365	0.18510	0.19294	0.19859
rec + ls + shuffle (u)	0.24584	0.12496	0.20950	0.25071	0.26183

Table 3. Coverage rates for confidence intervals of the gender pay gap in each educational sector between the original and perturbed data

Data	ISCED 0 and 1	ISCED 2	ISCED 3 and 4	ISCED 5A	ISCED 5B
rec + ls + ma	98.25	98.55	96.21	88.45	88.65
rec + ls + noise	89.85	99.86	87.81	91.58	99.26
pram + ma	83.52	96.63	83.45	78.18	95.08
rec + ls + shuffle	81.67	13.51	0.00	0.00	64.67

anonymisation methods. For those anonymisation methods that use (optimal) local suppression, three-anonymity is achieved. 4,414 observations violate three-anonymity in the original data set. The PRAM method performs best in terms of data utility since none of the variables that are used in these estimations are altered. The second best is recoding + local suppression + microaggregation. Recoding + local suppression + shuffling performs worst. The reasons for this could be that continuous variables are shuffled and also shuffled between gender, which is the most important variable when estimating the GPG and that the prediction quality of the model used for the shuffling procedure is low.

In general, recoding + local suppression + microaggregation and pram + microaggregation reports very low bias and clearly outperform shuffling and adding noise.

7.2. Overlap of Confidence Intervals

As an example, the upper and lower confidence intervals for the GPG in the domain *education* are given in Table 2. It is easy to see that the length of the confidence intervals is shorter for category ISCED 3-4 and largest for ISCED 0-1.

Again, the shuffling method does not seem to be able to give approximately the same confidence intervals.

A clearer picture is supported by Table 3, where the overlap of the confidence intervals for the GPG – estimated from the perturbed and the original data – is reported.

The coverage rates are relatively high for all methods except recoding + local suppression + shuffling. Differences in some categories are visible when comparing the other methods, whereas no clear ranking of them in terms of quality can be made.

The coverage rates for the gender pay gap in domain *age* (Table 4) are similar. Mostly the recoding + local suppression + microaggregation methods performs slightly better than recoding + local suppression + adding noise and pram + microaggregation.

However, a completely different picture is seen for the absolute relative bias of the Gini index in Table 5. Recoding + local suppression + microaggregation outperforms all other

Table 4. Coverage rates for confidence intervals of the GPG in each age class between the original and perturbed data

Data	(0,19)	(19,29)	(29,39)	(39,49)	(49,59)	(59,120)
rec + ls + ma	98.81	76.40	99.28	82.41	95.82	91.45
rec + ls + noise	94.90	80.27	94.31	89.60	89.70	96.76
pram + ma	84.26	88.92	95.02	88.55	92.58	86.94
rec + ls + shuffle	0.00	32.75	0.00	0.00	0.00	0.00

Table 5. Coverage rates for confidence intervals of the Gini indices in each age × gender domain between the original and perturbed data

Data	(0,19):f	(0,19):m	(19,29):f	(19,29):m	(29,39):f	(29,39):m
rec + ls + ma	93.64	81.66	96.83	94.93	89.24	95.63
rec + ls + noise	52.71	0.00	22.12	37.18	63.09	87.92
pram + ma	88.29	82.49	88.39	93.05	85.36	94.50
rec + ls + shuffle	82.61	0.00	0.00	0.00	20.38	0.00
	(39,49):f	(39,49):m	(49,59):f	(49,59):m	(59,120):f	(59,120):m
rec + ls + ma	84.69	75.33	99.21	94.59	95.40	92.22
rec + ls + noise	88.49	83.07	80.52	89.70	93.94	96.03
pram + ma	97.89	85.00	96.78	82.25	88.31	94.94
rec + ls + shuffle	12.55	55.93	0.00	0.00	0.00	0.00

Table 6. Regression coefficients

	Original	rec + ls + ma	rec + ls + noise	pram.ma	rec + ls + shuffle
(Intercept)	1.50454	1.52627	1.51374	1.40474	1.63726
Sexmale	0.20478	0.20484	0.20433	0.20970	0.19733
age(19,29]	0.57210	0.57190	0.58560	0.57659	0.76536
age(29,39]	0.73750	0.73745	0.75186	0.74388	0.91469
age(39,49]	0.81758	0.81746	0.83260	0.82634	0.96199
age(49,59]	0.85660	0.85597	0.87072	0.86754	0.89338
age(59,120]	0.81553	0.81067	0.82604	0.82169	0.49264
educationISCED 2	0.03692	0.02006	0.01011	0.03834	−0.25102
educationISCED 3 and 4	0.28314	0.26646	0.25737	0.28667	−0.16874
educationISCED 5A	0.73406	0.71508	0.70647	0.74198	0.09813
educationISCED 5B	0.44484	0.42802	0.41959	0.45337	−0.01251
LocationAT2	−0.07516	−0.07523	−0.07528	−0.06368	−0.00673
LocationAT3	−0.01230	−0.01207	−0.01132	−0.00900	−0.00098
NACE1D-Manufactoring	−0.05542	−0.06029	−0.05441	0.01740	−0.01600
NACE1E-Electricity	0.09709	0.09018	0.09264	0.12244	−0.02588
NACE1F-Construction	−0.12280	−0.12891	−0.12260	−0.03775	−0.01806
NACE1G-Trade	−0.18916	−0.19422	−0.18848	−0.09872	−0.02576
NACE1H-Hotels	−0.37478	−0.37962	−0.37589	−0.24398	−0.02269
NACE1I-Transport	−0.17130	−0.17632	−0.17061	−0.07943	−0.00939
NACE1J-FinancInt	0.14921	0.14532	0.15055	0.19273	−0.01993
NACE1K-RealEstate	−0.13433	−0.13901	−0.13517	−0.05156	−0.02072
NACE1M-Education	−0.16289	−0.16650	−0.16300	−0.07845	−0.02505
NACE1N-Health	−0.11299	−0.11734	−0.11360	−0.02939	−0.01838
NACE1O-Other	−0.19113	−0.19585	−0.19353	−0.10283	−0.01054

methods. PRAM + microaggregation also gives acceptable results but recoding + local suppression + adding noise gives low coverage rates for age classes below 29 years. Shuffling results in the estimates with the highest bias.

7.3. *Differences in Regression Coefficients*

As already mentioned, to compare the regression coefficients of original and anonymised data sets, the same categories in the explanatory variables of the model must be present. Thus the recoded twelve categories of *economic activity* are used also for the original data set, keeping in mind that this means a certain kind of information loss.

In Table 6 the regression coefficients for the original and the anonymised data sets are shown.

The regression coefficients and their confidence intervals are visualised in Figure 1, whereas the original estimates (in black) are compared with the estimates from anonymised data (in grey).

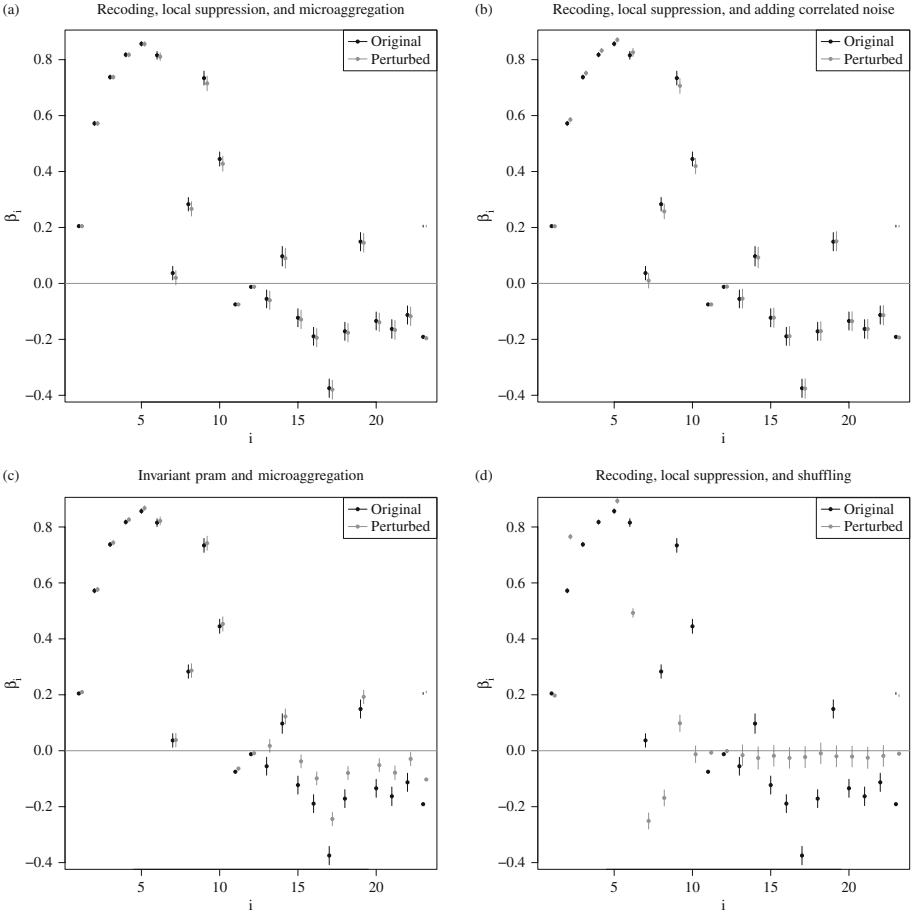

Fig. 1. Confidence intervals for the regression coefficients for the original data (black lines) and the perturbed data (grey dotted lines).

Recoding + local suppression + microaggregation again performs best and the confidence intervals obtained from the anonymised data almost always cover the confidence intervals obtained from the original data completely. Almost as good is the quality of data anonymised by recoding + local suppression + adding correlated noise. The results from invariant pram + microaggregation are good for all coefficients except those related to *economic activity*. This is not surprising, since this variable was one of the variables which was changed using PRAM. Some few coefficients are well preserved from the recoding + local suppression + shuffling anonymised data, but others are not. The reason is that even if the distribution of the continuous shuffled variables is well preserved, the relation to other variables that are not included in the shuffling model might be not preserved. A better model would probably lead to better results.

8. Conclusions

This article focuses upon the use of the most important measures of a particular survey as quality indicators of utility to evaluate anonymised data sets.

As a case study, the use of the Structure of Earnings Survey is analysed in detail in order to identify the most important variables, indicators and models applied to this data set. Based on the knowledge gained, the most important indicators are selected and the data utility of the anonymised data is evaluated; the disclosure risk is briefly reported. The evaluation is done on point and variance estimates from the selected indicators as well as on inferences on regression coefficients of a selected model. The evaluation of the regression coefficients in particular shows various problems with data utility. Thus such a comparison of model estimates should always be focused upon especially because a model reflects the multiple relationships between variables. Out of hundreds of different possible models, those models that are most often applied in practice should be chosen and an analysis of the literature is therefore necessary. The aim is to preserve the estimates from the most-used indicators and models and those anonymisations should be chosen that achieve both the minimum requirements in terms of disclosure risk and high precision on the chosen quality indicators.

The aim of this investigation was not to find the best anonymisation procedure from a risk perspective, but how to evaluate data utility. Nevertheless, four different possible anonymisations were applied and evaluated. The best results are obtained by the anonymisation: recoding + local suppression to achieve three-anonymity + microaggregation in each stratum defined by economic activity. For the invariant pram method, some problems become visible for those variables that have been 'pramed'. The shuffling method did not perform well, but this may depend on the shuffling model used (in our study several models were tested and the best was chosen); good results on other data sets may perform better as the method seems very promising (see, e.g., Muralidhar and Sarathy 2006).

This case study is only focused on one particular survey, the Structural Earnings Statistics survey, but we have demonstrated a general concept of how to identify the most important indicators and models and how to evaluate the quality of the protected data based on estimates of these indicators. Although this key idea is not new in priciple, it is demonstrated practically in a large case study in a larger setting.

The used (and other) indicators have been implemented in the R package **laeken** (Alfons and Templ 2013), which makes the application of the methods to complex data, such as the SES, easy.

9. References

Alfons, A. and M. Templ. 2013. "Estimation of Social Exclusion Indicators from Complex Surveys: The R package laeken." *Journal of Statistical Software* 54: 1–25.

Alfons, A., S. Kraft, M. Templ, and P. Filzmoser. 2011. "Simulation of Close-to-Reality Population Data for Household Surveys with Application to EU-SILC." *Statistical Methods & Applications* 20: 383–407. doi:10.1007/s10260-011-0163-2.

Alfons, A., M. Templ, and P. Filzmoser. 2013. "Robust Estimation of Economic Indicators from Survey Samples Based on Pareto Tail Modeling." *Journal of the Royal Statistical Society Series C* 62: 271–286.

Beblot, M., D. Beniger, A. Heinze, and F. Laisney. 2003. *Methodological Issues Related to the Analysis of Gender Gaps in Employment, Earnings and Career Progression*. Final Project Report, European Commission Employment and Social Affairs DG.

Belfield, R. 1999. *Pay Inequalities and Economic Performance: A Review of the UK Literature*. Technical Report PiEP Report, Centre for Economic Performance, London School of Economics.

Bowles, S., H. Gintis, and M. Osborne. 2001. "The Determinants of Earnings: a Behavioral Approach." *Journal of Economic Literature* 39: 1137–1176.

Brand, R. 2004. "Microdata Protection through Noise Addition." In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 347–359. New York: Springer.

Bruch, C., R. Münnich, and S. Zins. 2011. *Variance Estimation For Complex Surveys*. Research Project Report WP3–D3.1, FP7-SSH-2007-217322 AMELI. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Caju, P., C. Fuss, and L. Wintr. 2009a. "Understanding Sectoral Differences in Downward Real Wage Rigidity: Workforce Composition, Institutions, Technology and Competition." Working Paper Series no. 1006, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1006.pdf (accessed December 2013)

Caju, P., F. Rycx, and I. Tojerow. 2009b. "Inter-industry Wage Differentials: How Much Does Rent Sharing Matter?" *Journal of the European Economic Association* 79: 691–717.

Caju, P., F. Rycx, and I. Tojerow. 2010. "Wage Structure Effects of International Trade: Evidence From a Small Open Economy." Working Paper Series no. 1325, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1325.pdf (accessed December 2013)

Carlson, M. 2002. "Assessing Microdata Disclosure Risk Using the Poisson-inverse Gaussian Distribution." *Statistics in Transition* 5: 901–925.

Casali, S. and V. Alvarez. 2010. *17% of Full-time Employees In the EU Are Low-wage Earners. Statistics in focus*. Research Report. KS-SF-10-003-EN-N, Eurostat/European Commission. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-SF-10-003/EN/KS-SF-10-003-EN.PDF (accessed December 2013)

Dell'Aringa, C., P. Ghinetti, and C. Lucifora. 2000. "Pay Inequality and Economic Performance in Italy: a Review of the Applied Literature." In Proceedings of the LSE conference, November 3–4, 2000. 1–28. London.

Deville, J.-C. and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87: 376–382.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–1020.

Domingo-Ferrer, J. and V. Torra. 2001. "A Quantitative Comparison of Disclosure Control Methods for Microdata." *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, edited by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz. 111–134, Eurostat.

Domingo-Ferrer, J., J.M. Mateo-Sanz, and T. Torra. 2001. "Comparing sdc Methods for Microdata on the Basis of Information Loss and Disclosure." Proceedings of ETK-NTTS 2001: Eurostat, Luxembourg June 18–20, 2001. 807–826. Luxembourg: Eurostat.

Dupray, D., H. Nohara, and P. Béret. 1999. *Pay Inequality and Economic Performance: a Review of the French Literature*. Technical Report PiEP Report, Centre for Economic Performance, London School of Economics

Dupré, D. 2010. "The Unadjusted Gender Pay Gap in the European Union." In Joint UNECE/Eurostat Work Session on Gender Statistics, Geneva April 14–16, 2010. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.30/2010/1.e.pdf (accessed November 2015)

Dybczak, K., and K. Galuscak. 2010. "Changes in the Czech Wage Structure: Does Immigration Matter?" Working Paper Series no. 1242, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1242.pdf (accessed December 2013)

Edwards, C. 2010. "Public Sector Unions and the Rising Costs of Employee Compensation," *Cato Journal* 30: 87–115.

Efron, B. and R.J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

EU-SILC. 2004. *Common Cross-sectional EU Indicators Based on EU-SILC: the Gender Pay Gap*. EU-SILC 131-rev/04, Working Group on Statistics on Income and Living Conditions (EU-SILC). Luxembourg: Eurostat.

EU-SILC. 2009. Algorithms to Compute Social Inclusion Indicators Based On EU-SILC and Adopted under the Open Method of Coordination (OMC). EU-SILC LC-ILC/39/09/ENrev.1, Directorate F: Social and Information Society Statistics Unit F-3: Living Conditions and Social Protection, European Commission. Luxembourg: Eurostat.

Fitzenberger, B., K. Kohn, and A. Lembcke. 2006. *Union Wage Effects in Germany: Union Density Or Collective Bargaining Coverage?* Research Report FSP 1169, DFG research programme, The London School of Economics and Political Sciences, London.

Franconi, L. and S. Polettini. 2004. "Individual Risk Estimation in μ-Argus: a Review." In *Privacy in Statistical Databases: Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 262–272. New York: Springer.

Franconi, L., D. Ichim, and M. Templ. 2011. *First Steps to Define a Framework For Comparable Dissemination of the European Structure of Earning Survey. Deliverable d1.1-a. Task 1: Harmonisation of Microdata Release in Multiple Countries*. Essnet Project on Common Tools and Harmonised Methodologies for SDC in the ESS. Available at: http://neon.vb.cbs.nl/casc/..%5Ccas%5CESSNet2%5Cdeliverable%201%20full%20august2012.pdf (accessed November 2015)

Frick, B., and K. Winkelmann. 1999. *Pay Inequalities and Economic Performance: A Review in Literature*, Technical Report Research Report HPSE-CT-1999-00040, Ernst-Moritz-Arndt-Universität Greifswald.

Geissberger, T. 2009. *Verdienststrukturerhebung 2006, Struktur und Verteilung der Verdienste in Oösterreich*. Vienna: Statistik Austria.

Geissberger, T. 2010. *Frauenbericht. Teil 4: Sozioökonomische Studien*, Technical Report 4, Federal Ministry for Women and the Civil Service of Austria.

Geissberger, T. and K. Knittler. 2010. "Niedriglöhne und Atypische Beschäftigung in Österreich." *Statistische Nachrichten* 6: 448–461.

Gini, C. 2012. "Variabilità e Mutabilità: Contributo Allo Studio delle Distribuzioni e delle Relazioni Statistiche." *Studi Economico-Giuridici della R. Università di Cagliari* 3: 3–159.

Gomatam, S. and A. Karr. 2003. *Distortion Measures for Categorical Data Swapping*. Report no. 131, National Institute of Statistical Sciences (NISS).

Gouweleeuw, J., P. Kooiman, L. Willenborg, and P-P. De Wolf. 1998. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." *Journal of Official Statistics* 14; 463–478.

Graf, M., A. Alfons, C. Bruch, P. Filzmoser, B. Hulliger, R. Lehtonen, B. Meindl, R. Münnich, T. Schoch, M. Templ, M. Valaste, A. Wenger, and S. Zins. 2011. *State-of-the-art of laeken Indicators*. Research Project Report WP1 – D1.1, FP7-SSH-2007-217322 AMELI. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Groshen, E. 1991. "The Structure of the Female/Male Wage Differential." *Journal of Human Resources* 26: 455–472.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.-P. de Wolf. 2012. *Statistical Disclosure Control*. New York: Wiley.

Ichim, D. and L. Franconi. 2007. "Disclosure Scenario and Risk Assessment: Structure of Earnings Survey." In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Manchester, December 17–19, 2007. Doi: 10.2901/Eurostat.C2007.004

Ichim, D. and L. Franconi. 2010. "Strategies to Achieve sdc Harmonisation at European Level: Multiple Countries, Multiple Files, Multiple Surveys." *Privacy in Statistical Databases '10*, edited by J. Domingo-Ferrer and E. Kajkos, Springer, New York. 284–296.

Karr, A.F., C.N. Kohnen, A. Oganian, J.P. Reiter, and A.P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60: 224–232. Doi: 10.1198/000313006X124640.

Kolb, J.-P., R. Münnich, S. Beil, A. Chatziparadeisis, and J. Seger. 2011. *Policy Use of Indicators on Poverty and Social Exclusion*. Research Project Report WP9–D9.1,

FP7-SSH-2007-217322 AMELI, 2011. Available at: http://ameli.surveystatistics.net (accessed December 2013)

Lorenz, M.O. 1905. "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association* 9: 209–219.

Manning, A.M., D.J. Haglin, and J.A. Keane. 2008. "A Recursive Search Algorithm For Statistical Disclosure Assessment." *Data Mining and Knowledge Discovery* 16: 165–196. Doi: 10.1007/s10618-007-0078-6.

Marsden, D. 2010. *Pay Inequalities and Economic Performance*, Technical Report PiEP Final Report V4, Centre for Economic Performance, London School of Economics. London: London School of Economics. Available at: http://www.ist-world.org/ProjectDetails.aspx?ProjectID=fa5bb4adfff74d60aeca90b56441a601&SourceDatabaseID=9cd97ac2e51045e39c2ad6b86dcelac2.

Messina, J., M. Izquierdo, P. Caju, C.F. Duarte, and N.L. Hanson. 2010. "The Incidence of Nominal and Real Wage Rigidity: an Individual-based Sectoral Approach." *Journal of the European Economic Association* 8: 487–496.

Mittag, J. 2005. *Gross Earnings In Europe. Main Results of the Structure of Earnings Survey 2002*. Statistics in Focus. Research Report. KS-NK-05-012-EN-N, European Communities. Available at: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-NK-05-012/EN/KS-NK-05-012-EN.PDF (accessed December 2013)

Muralidhar, K. and R. Sarathy. 2006. "Data Shuffling – a New Masking Approach for Numerical Data." *Management Science* 52: 658–670.

Nolan, B. and H. Russel. 2001. *Pay Inequality and Economic Performance In Ireland: a Review of the Applied Literature*. Technical Report PiEP Report, The Economic and Social Research Institute, Dublin.

Oganian, A. and A.F. Karr. 2006. "Combinations of sdc Methods for Microdata Protection." In *Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and L. Franconi. 102–113. Berlin: Springer. Doi: 10.1007/11930242_10.

Pointner, W., and A. Stiglbauer. 2010. "Changes In the Austrian Structure of Wages." Working Paper Series no. 1268, European Central Bank. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1268.pdf (accessed December 2013)

Reiter, J.P. 2012. "Statistical Approaches to Protecting Confidentiality For Microdata and their Effects on the Quality of Statistical Inferences." *Public Opinion Quarterly* 76: 163–181. Doi: 10.1093/poq/nfr058.

Research Center for Education and the Labour Market at Maastricht University. 2009. "Development of Econometric Methods to Evaluate the Gender Pay Gap Using Structure of Earnings Survey Data." Research paper no. ks-ra-09-011-en-n, European Commission. Available at: http://www.ecb.int/pub/pdf/scpwps/ecbwp1006.pdf (accessed December 2013)

Reuter, W. 2010. *Establishing an Infrastructure for Remote Access to Microdata at Eurostat*. Bachelor's thesis., Vienna Univesity of Economics.

Reuter, W. and J-M. Museux 2010. "Establishing an Infrastructure for Remote Access to Microdata at Eurostat." In *Privacy in Statistical Databases: Lecture Notes in Computer Science*, edited by J. Domingo-Ferrer. 249–257. New York: Springer.

Rinott, Y. 2003. "On Models for Statistical Disclosure Risk Estimation." In Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. April 7–9, 2003. 275–285, United Nations Statistical Commission, Geneva.

Shlomo, N. 2008. "Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Data Utility." In Section on Survey Research Methods, JSM. August 3–7, 2008, Denver, Colorado, USA. 229–240. Available at: https://www.amstat.org/sections/srms/proceedings/y2008/Files/300242.pdf (accessed November 2015)

Simón, H. 2010. "International Differences in Wage Inequality: A New Glance with European Matched Employer-Employee Data." *British Journal of Industrial Relations* 48: 310–346.

Stephan, G. and K. Gerlach. 2005. "Wage Settlements and Wage Settings: Evidence from a Multilevel Model." *Applied Economics* 37: 2297–2306.

Stockinger, S. 2010. *Frauenbericht 2010*. Technical report, Federal Ministry for Women and the Civil Service of Austria. Vienna: Available at: http://www.bka.gv.at/site/6811/default.aspx (accessed December 2013)

Sweeney, L. 2002. "k-Anonymity: a Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10: 557–570.

Templ, M. 2008. "Statistical Disclosure Control for Microdata Using the R-package sdcMicro." *Transactions on Data Privacy* 1: 67–85.

Templ, M. 2011a. *Estimators and Model Predictions from the Structural Earnings Survey for Benchmarking Statistical Disclosure Methods*. Research Report CS-2011-4, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria.

Templ, M. 2011b. "Comparison of Perturbation Methods Based on Pre-defined Quality Indicators." In Joint UNECE/Eurostat work session on statistical data confidentiality, 26–28 October, 2011, Tarragona, Spain, 1–10. Unece, Geneva, Italy.

Templ, M. and A. Alfons. 2011. *Variance Estimation of Social Inclusion Indicators Using the R Package laeken*. Research Report CS-2011-3, Department of Statistics and Probability Theory, Vienna University of Technology. Available at: http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-3complete.pdf (accessed December 2013)

Templ, M. and P. Filzmoser. 2014. "Simulation and Quality of a Synthetic Close-to-Reality Employer-Employee Population." *Journal of Applied Statistics*, 41: 1053–1072.

Templ, M. and B. Meindl. 2010. "Practical Applications in Statistical Disclosure Control Using R." In *Privacy and Anonymity in Information Management Systems: Advanced Information and Knowledge Processing*, edited by J. Nin and J. Herranz. 31–62. London: Springer.

Templ, M. A. Kowarik, and B. Meindl. 2015. "Statistical Disclosure Control for Micro-Data Using R Package sdcMicro." *Journal of Statistical Software*. 67: 1–36.

Weinberg, D.H. 2007. "Earnings by Gender: Evidence from Census 2000." *Monthly Labor Review Online* 130: 26–34.

Winter-Ebmer, R. and J. Zweimüller. 1999. "Firm Size Wage Differentials in Switzerland: Evidence from Job Changers." *American Economic Review* 89: 89–93.

Woo, M., J.P. Reiter, A. Oganian, and A.F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1: 111–124.

Received October 2012
Revised December 2013
Accepted January 2015