Disclosure Risk of Synthetic Population Data with Application in the Case of EU-SILC*

Matthias Templ[1,2] and Andreas Alfons[1]

[1] Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstraße 7, 1040 Vienna, Austria
templ@tuwien.ac.at, alfons@statistik.tuwien.ac.at
[2] Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

Abstract. In survey statistics, simulation studies are usually performed by repeatedly drawing samples from population data. Furthermore, population data may be used in courses on survey statistics to support the theory by practical examples. However, real population data containing the information of interest are in general not available, therefore synthetic data need to be generated. Ensuring data confidentiality is thereby absolutely essential, while the simulated data should be as realistic as possible. This paper briefly outlines a recently proposed method for generating close-to-reality population data for complex (household) surveys, which is applied to generate a population for Austrian EU-SILC (European Union Statistics on Income and Living Conditions) data. Based on this synthetic population, confidentiality issues are discussed using five different worst case scenarios. In all scenarios, the intruder has the complete information on key variables from the real survey data. It is shown that even in these worst case scenarios the synthetic population data are confidential. In addition, the synthetic data are of high quality.

Keywords: Survey Statistics, Synthetic Population Data, Data Confidentiality

1 Introduction

In the analysis of survey data, variability due to sampling, imputation of missing values, measurement errors and editing must be considered. Statistical methods thus need to be evaluated with respect to the effect of these variabilites on point and variance estimates. A frequently used strategy to adequately measure such effects under different settings is to perform simulation studies by repeatedly drawing samples from population data (possibly using different sampling designs) and to compare the estimates with the true values of the sample frame.

* This work was partly funded by the European Union (represented by the European Commission) within the 7th framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). For more information on the project, visit http://ameli.surveystatistics.net.

Evaluating and comparing various statistical methods within such a *design-based* simulation approach under different *close-to-reality* settings is daily work for survey statisticians and has been done, e.g., in the research projects DACSEIS [1], EurEdit [2] and AMELI [3].

Furthermore, population data may be used for teaching courses on survey statistics. Realistic examples may help students to better understand issues in survey sampling, e.g., regarding different sampling designs.

Since suitable population data are typically not available, it is necessary to generate synthetic data. The generation of population microdata for selected surveys as a basis for Monte Carlo simulation studies is described in [1, 4]. These procedures were extended in [3, 5] to simulate close-to-reality population data for more complex surveys such as EU-SILC (*European Union Statistics on Income and Living Conditions*). However, confidentiality issues of such synthetic population data are only briefly addressed in these contributions.

Generation of population microdata for simulation studies is closely related to the field of *microsimulation* [6], yet the aims are quite different. Microsimulation models attempt to reproduce the behavior of individual units within a population for policy analysis purposes and are well-established within the social sciences. Nevertheless, they are highly complex and time-consuming. On the other hand, synthetic population microdata for simulation studies in survey statistics are used to evaluate the behavior of statistical methods. Thus fast computation is preferred to over-complex models.

Another approach towards the generation of microdata is to use multiple imputation to create *fully* or *partially* synthetic data sets, as proposed in [7, 8]. This approach is further discussed in [9–11]. However, these contributions do not consider some important issues such as the generation of categories that do not occur in the original sample or the generation of structural zeros.

The rest of the paper is organized as follows. Section 2 outlines the framework for generating synthetic populations proposed in [5]. Then the data investigated in this paper are introduced in Section 3. Sections 4 and 5 discuss statistical disclosure control issues related to survey and population data. In Section 6, several scenarios for evaluating the confidentiality of synthetic population data are described, while Section 7 lists the obtained results for these scenarios. The final Section 8 concludes.

2 Generation of Synthetic Population Data

The generation of synthetic population data for surveys is described in great detail in [5]. Therefore, only the basic ideas of this framework are presented here. Several conditions for simulating population data are listed in [1, 4, 5]. The most important requirements are:

- Actual sizes of regions and strata need to be reflected.
- Marginal distributions and interactions between variables should be represented accurately.

- Heterogeneities between subgroups, in particular regional aspects, should be allowed.
- Data confidentiality must be ensured.

In general, the framework for generating synthetic population data consists of four steps:

1. In case of household data, set up the household structure.
2. Simulate categorical variables.
3. Simulate continuous variables.
4. Split continuous variables into components.

Not all of these steps need to be performed, depending on the survey of interest.

Step 1. When generating household data, the household structure is simulated separately for the different household sizes within each strata. Using the corresponding sample weights, the number of households is simply estimated by the Horvitz-Thompson estimator [12]. The structure of the population households is then simulated by resampling some basic variables from the sample households with probabilities proportional to the sample weights. For disclosure reasons, information from as few variables as possible should be used to construct the household structure (e.g., only age and gender information).

Step 2. For each stratum, the conditional distribution of any additional categorical variable is estimated with a multinomial logistic regression model. The previously simulated variables are thereby used as predictors. Furthermore, the sample weights are considered and it is possible to account for structural zeros. The main advantage of this approach is that it allows to generate combinations that do not occur in the sample, which is not the case for the procedure introduced in [1, 4].

Step 3. Two approaches for simulating continuous variables are proposed in [5], but only the approach that performs better in the case of EU-SILC data is considered in this paper. First, the variable to be simulated is discretized using suitable breakpoints. The discretized variable is then then simulated as described in the previous step. Finally, the values of the continuous variable are randomly drawn from uniform distributions within the respective intervals. Note that the idea behind this approach is to divide the data into relatively small subsets so that the uniform distribution is not too much of an oversimplification.

Step 4. Splitting continuous variables into components is based on conditional resampling of fractions from the sample households with probabilities proportional to the sample weights. Only very few highly influential categorical variables should thereby be considered for conditioning. The resampled fractions are then multiplied with the previously simulated total.

The data simulation framework proposed in [5] is implemented in the R [13] package simPopulation [14]. In addition to the four steps of the procedure and a wrapper function to generate synthetic EU-SILC populations, various diagnostic plots are available.

3 Synthetic EU-SILC Population Data

The *European Union Statistics on Income and Living Conditions* (EU-SILC) is a complex panel survey conducted in EU member states and other European countries. It is mainly used for measuring risk-of-poverty and social cohesion in Europe [15]. The generation of synthetic population data based on Austrian EU-SILC survey data from 2006 is discussed and evaluated in [5]. The resulting synthetic population is investigated in this paper with respect to confidentiality issues. Table 1 lists the variables that are used in the analysis. A detailed description of all variables included in EU-SILC data and their possible outcomes is given in [16].

Table 1. Variables of the synthetic EU-SILC population data used in this paper.

Variable	**Type**	
Region	Categorical	9 levels
Household size	Categorical	9 levels
Age category	Categorical	15 levels
Gender	Categorical	2 levels
Economic status	Categorical	7 levels
Citizenship	Categorical	3 levels
Personal net income	Semi-continuous	

In order to demonstrate that the synthetic population data are of high quality, they are compared to the underlying sample data. Figure 1 contains mosaic plots of gender, region and household size for the sample and synthetic population data, respectively. Clearly, the plots show almost identical structures. In addition, the distribution of personal net income is visualized in Figure 2. On the left hand side, the cumulative distribution functions for the sample and population data, respectively, are displayed. For better visibility, only the main parts of the data are shown, which are nearly in perfect superposition. On the right hand side, the conditional distributions with respect to gender are represented by box plots. The fit of the distribution within the subgroups is excellent and heterogeneities between the subgroups are very well reflected. Note that points outside the extremes of the whiskers are not plotted. For extensive collections of results showing that the multivariate structure of the data is well preserved, the reader is referred to [5, 17].

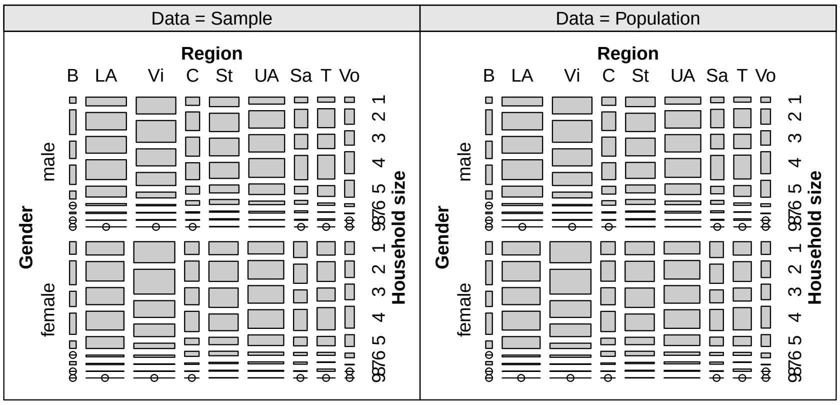

Fig. 1. Mosaic plots of gender, region and household size of the Austrian EU-SILC sample from 2006 and the resulting synthetic population.

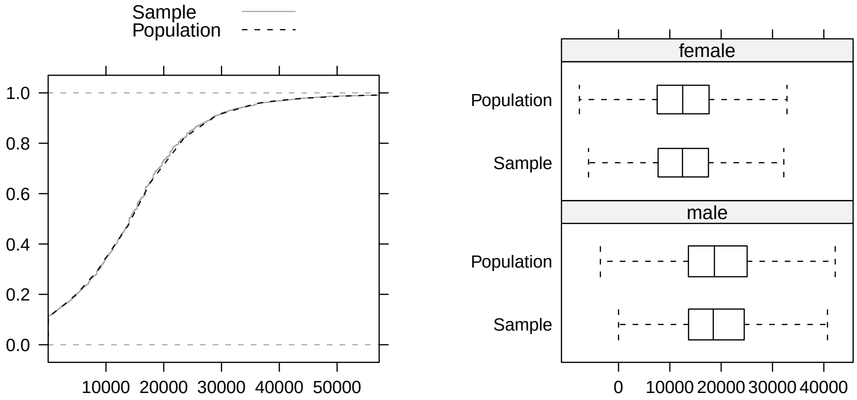

Fig. 2. Personal net income in the Austrian EU-SILC sample from 2006 and the resulting synthetic population. *Left*: Cumulative distribution functions of personal net income. Only the main parts of the data are shown for better visibility. *Right:* Box plots of the conditional distributions with respect to gender. Points outside the extremes of the whiskers are not plotted.

4 A Global Disclosure Risk Measure for Survey Data

A popular global measure of the reidentification risk for survey data is given by the number of uniquenesses in the sample that are unique in the population as well. Let m categorical key variables in the sample and population data be denoted by $\boldsymbol{x}_j^S = (x_{1j}^S, \dots, x_{nj}^S)'$ and $\boldsymbol{x}_j^P = (x_{1j}^P, \dots, x_{Nj}^P)'$, respectively, $j = 1, \dots, m$, where n and N give the corresponding number of observations. For an observation in the sample given by the index $c = 1, \dots, n$, let J_c^S and J_c^P denote the index sets of observations in the sample and population data, respectively, with equal values in the m key variables:

$$\begin{aligned} J_c^S &:= \{j = 1, \dots, n : x_{jk}^S = x_{ck}^S,\ k = 1, \dots, m\}, \\ J_c^P &:= \{j = 1, \dots, N : x_{jk}^P = x_{ck}^S,\ k = 1, \dots, m\}. \end{aligned} \tag{1}$$

Furthermore, a function $\mathcal{I}$ is defined as

$$\mathcal{I}(J) := \begin{cases} 1 & \text{if } |J| = 1, \\ 0 & \text{else.} \end{cases} \tag{2}$$

The global disclosure risk measure can then be expressed by

$$\tau_0 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^P). \tag{3}$$

Note that the notation in (3) differs from the common definition. For comparison, see, e.g., the risk measures in [18, 19]. The notation used in (3) describes the same phenomenon, but provides more flexibility in terms of software implementation [20] and allows to formulate the adapted risk measures given in the following section.

Clearly, the risk of reidentification is lower the higher the corresponding population frequency count. If the population frequency count is sufficiently high, it is not possible for an intruder to assign the observation for which they hold information with absolute certainty. Hence the intruder does not know whether the reidentification was successful. However, the true frequency counts of the population are usually unknown and need to be estimated by modeling the distribution of the population frequencies.

In Section 6, the global disclosure risk measure τ_0 is modified to estimate the disclosure risk for synthetic population data in certain scenarios instead of survey data.

5 Confidentiality of Synthetic Population Data

The motivation for generating close-to-reality population data is to make the resulting data sets publicly available to researchers for use in simulation studies or courses on survey statistics. Therefore, the disclosure risk of such data needs to be low, while at the same time the multivariate structure should be as realistic as possible.

If population data are generated from perturbed survey data, confidentiality is guaranteed whenever the underlying survey data are confidential. Perturbing survey data is typically done by performing recodings and suppression such that k-anonymity [21, 22] is provided for categorical key variables, as well as low risk of reidentification on the individual level is ensured [23–25, and references therein]. In any case, perturbation implies information loss. Usually not all combinations of categorical key variables are still included in the perturbed sample and outliers in continuous variables are often modified to a great extent. It is thus favorable to use information of the unperturbed sample to generate synthetic populations, as this increases the quality of the resulting data.

Based on the ideas proposed in [7, 8], the generation of *fully* or *partially* synthetic population data using multiple imputation is discussed in [9–11]. More precisely, let p be the number of variables in the sample and let the first k, $1 \leq k < p$, categorical variables be available for the population data from administrative sources. These first k variables are released unchanged, while the remaining $p - k$ variables are estimated using regression based multiple imputation. It is important to note that the first k variables of real population data may still contain unique combinations after cross tabulation, therefore further investigation may be necessary to ensure confidentiality. Probabilities of reidentification for such synthetic data have been studied in [26], based on the work of [27, 28], by matching the synthetic data with the intruder's data on some predefined key variables.

The situation for synthetic population data generated by the approach of [5] is somewhat different. A very low number of basic categorical variables are generated in the first step by resampling from the actual survey data. Since the sample weights are thereby used as probability weights, on average k-anonymity is provided with respect to these basic variables, where k denotes the smallest sample weight. In surveys, k is typically very high (> 500), hence the disclosure risk is very low. However, additional categorical and continuous variables are generated based on models obtained from the actual survey data. In particular, the generation of continuous variables involves random draws from certain intervals.

With the additional categorical variables, some unique combinations may be introduced in the synthetic population data. If such a combination is not unique in the real population, it is not useful to an intruder. On the other hand, if such a combination is unique in the real population as well, it must be ensured that the values of the other variables in the synthetic population data are not too close to the real values. Most notably, it is of interest to measure the difference in continuous variables of the successfully matched statistical units.

In addition, unique combinations in the real population may even be critical if they are not unique in the synthetic population data. An intruder could in this case look for all occurrences of such a combination in the synthetic population. If the corresponding units have too similar values in a (continuous) variable of interest, the intruder may be able to infer some information on the original value, since the synthetic values have been predicted with models obtained from the real sample data.

In order to investigate these issues in more detail, various disclosure scenarios are introduced in the following section. Section 7 then presents the results for the synthetic EU-SILC population data described in Section 3.

6 Disclosure Scenarios for Synthetic Population Data

Five different scenarios are considered to evaluate the confidentiality of synthetic data generated with the framework proposed in [5]. These scenarios are motivated by the synthetic EU-SILC population data, hence only a continuous variable is considered to contain confidential information, while there are m categorical key variables. In the case of EU-SILC, the confidential variable is *personal net income* and the key variables are *region*, *household size*, *age category*, *gender*, *economic status* and *citizenship* (see Table 1). Let the confidential continuous variable for the original sample and synthetic population, respectively, be denoted by $\boldsymbol{y}^S = (y_1^S, \ldots, y_n^S)'$ and $\boldsymbol{y}^U = (y_1^U, \ldots, y_N^U)'$, while the categorical key variables are denoted by $\boldsymbol{x}_j^S = (x_{1j}^S, \ldots, x_{nj}^S)'$ and $\boldsymbol{x}_j^U = (x_{1j}^U, \ldots, x_{Nj}^U)'$, $j = 1, \ldots, m$, analogous to the definitions in Section 4. Furthermore, let J_c^S be defined as in (1) and let J_c^U be defined accordingly as

$$J_c^U := \{j = 1, \ldots, N : x_{jk}^U = x_{ck}^S,\ k = 1, \ldots, m\}. \tag{4}$$

In the following scenarios, the intruder has knowledge of the m key variables for all observations from the original sample and tries to acquire information on the confidential variable.

It should be noted that the link to the global risk measure from (3) is loosened in the following. Disclosure is considered to occur if the value of the confidential variable for a unique combination of key variables in the sample can be closely estimated from the synthetic population data, given a prespecified value of accuracy p. However, such a sample uniqueness does not need to be unique in the true population, in which case close estimation of the confidential variable would not necessarily result in disclosure. In this sense, the following scenarios can be considered worst case scenarios and the reidentification risk is thus overestimated. Proper analysis with estimation of the true population uniquenesses is future work.

6.1 Scenario 1: Attack Using One-to-One Matches in Key Variables with Information on the Data Generation Process

The intruder in this scenario tries to find one-to-one matches between their data and the synthetic population data. Moreover, they know the intervals from which the synthetic values were drawn. For details on the data generation procedure, the reader is referred to [5]. Let these intervals be denoted by $[l_j, u_j]$, $j = 1, \ldots, N$, and let l be a function giving the length of an interval defined as

$l([a,b]) := b - a$ and $l(\emptyset) := 0$. With a prespecified value of accuracy p defining a symmetric interval around a confidential value, (3) is reformulated as

$$\tau_1 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^U) \cdot \frac{l([y_c^S(1-p), y_c^S(1+p)] \cap [l_{j_c}, u_{j_c}])}{l([l_{j_c}, u_{j_c}])}, \tag{5}$$

where $j_c \in J_c^U$ if $|J_c^U| = 1$, i.e., j_c is the index of the unit in the synthetic population with the same values in the key variables as the cth unit in the intruder's data if such a one-to-one match exists, otherwise it is a dummy index. The last term in (5) thereby gives the probability that for the successfully matched unit, the synthetic value drawn from the interval $[l_{j_c}, u_{j_c}]$ is sufficiently close to the original value y_c^S.

6.2 Scenario 2: Attack Using One-to-One Matches in Key Variables without Information on the Data Generation Process

In general, an intruder does not have any knowledge on the intervals from which the synthetic values were drawn. In this case, reidentification is successful if the synthetic value itself of a successfully matched unit is sufficiently close to the original value. The risk of reidentification thus needs to be reformulated as

$$\tau_2 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathcal{I}(J_c^U) \cdot \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(y_{j_c}^U), \tag{6}$$

where j_c is defined as above and $\mathbb{I}_A$ denotes the indicator function for a set A.

6.3 Scenario 3: Attack Using All Occurrences in Key Variables with Information on the Data Generation Process

This scenario is an extension of Scenario 1, in which the intruder does not only try to find one-to-one matches, but looks for all occurrences of a unique combination from their data in the synthetic population data. Keep in mind that the intruder in this scenario knows the intervals from which the synthetic values were drawn. For a unique combination in the intruder's data, reidentification is possible if the probability that the synthetic values of all matched units are sufficiently close to the original value. Hence the disclosure risk from (5) changes to

$$\tau_3 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \prod_{j \in J_c^U} \frac{l([y_c^S(1-p), y_c^S(1+p)] \cap [l_j, u_j])}{l([l_j, u_j])}. \tag{7}$$

6.4 Scenario 4: Attack Using All Occurrences in Key Variables without Information on the Data Generation Process

In an analogous extension of Scenario 2, reidentification of a unique combination from the intruder's data is successful if the synthetic values themselves of all matched units are sufficiently close to the original value. Equation (6) is in this case rewritten as

$$\tau_4 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \prod_{j \in J_c^U} \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(y_j^U). \tag{8}$$

6.5 Scenario 5: Attack Using Key Variables for Model Predictions

In this scenario, the intruder uses the information from the synthetic population data to obtain a linear model for $\boldsymbol{y}^U$ with predictors $\boldsymbol{x}_j^U$, $j = 1, \ldots, m$:

$$\boldsymbol{y}^U = \beta_0 + \beta_1 \boldsymbol{x}_1^U + \ldots + \beta_m \boldsymbol{x}_m^U + \boldsymbol{\varepsilon}. \tag{9}$$

For a unique combination of the key variables, reidentification is possible if the corresponding predicted value is sufficiently close to the original value. Let the predicted values of the intruder's data be denoted by $\hat{\boldsymbol{y}}^S = (\hat{y}_1^S, \ldots, \hat{y}_n^S)'$. Then the disclosure risk can be formulated as

$$\tau_5 := \sum_{c=1}^{n} \mathcal{I}(J_c^S) \cdot \mathbb{I}_{[y_c^S(1-p), y_c^S(1+p)]}(\hat{y}_c^S). \tag{10}$$

Note that for large population data, the computational costs for fitting such a regression model are very high, so an intruder needs to have a powerful computer with very large memory. Furthermore, the intruder could also perform a stepwise model search using an optimality criterion such as the AIC [29].

7 Results

The disclosure risk of the synthetic Austrian EU-SILC population data described in Section 3 is analyzed in the following with respect to the scenarios defined in the previous section. In these scenarios, the intruder has knowledge of the categorical key variables *region*, *household size*, *age category*, *gender*, *economic status* and *citizenship* for all observations in the original sample used to generate the data. In addition, the intruder tries to obtain information on the confidential variable *personal net income* (see Table 1 for a description of these variables). The original sample thereby consists of $n = 14\,883$ and the synthetic population of $N = 8\,182\,218$ observations.

Note that this paper only evaluates the risk of reidentification for this specific synthetic data set. In order to get more general results regarding confidentiality of the data generation process, many data sets need to be generated in a simulation study and the average values need to be reported. This task, however, is beyond the scope of this paper.

Table 2 lists the results for the risk measures for the investigated scenarios using different values of the accuracy parameter p. Besides the absolute values, the relative values with respect to the size of the intruder's data set are presented, which give the probabilities of successful reidentification.

Table 2. Results for Scenarios 1-5 using different values for the accuracy parameter p.

		$\boldsymbol{p}$		
Scenario	**Risk measure**	0.01	0.02	0.05
1	τ_1	0	0	0.052
2	τ_2	0	0	0
3	τ_3	$1.1 \cdot 10^{-8}$	$1.2 \cdot 10^{-6}$	0.053
4	τ_4	15	15	15
5	τ_5	20	43	110
1	τ_1/n	0	0	$3.5 \cdot 10^{-6}$
2	τ_2/n	0	0	0
3	τ_3/n	$6.7 \cdot 10^{-13}$	$8.6 \cdot 10^{-11}$	$3.5 \cdot 10^{-6}$
4	τ_4/n	0.001	0.001	0.001
5	τ_5/n	0.001	0.003	0.007

The results show that even if an intruder is able to reidentify an observation, they do not gain any useful information, as the probability that the obtained value is sufficiently close to the original value is extremely low.

In particular if the intruder tries to find one-to-one matches (Scenarios 1 and 2), the probability of a successful reidentification is only positive for $p = 0.05$ and if they have information on the data generation process, i.e., the intervals from which the synthetic values were drawn.

If the intruder looks for all occurrences of a unique combination from their data in the synthetic population, using information on the data generation process hardly changes the probabilities of reidentification (Scenario 3). This is not a surprise given the formula in (7), since for such a unique combination, the probabilities that the corresponding synthetic values are sufficiently close to the original value need to be multiplied. On the other hand, if the intruder uses only the synthetic values (Scenario 4), some observations are successfully reidentified. Nevertheless, the probabilities of reidentification are extremely low.

Among the considered scenarios, Scenario 5 leads to the highest disclosure risk. However, the regression model in this scenario comes with high computational costs and the probabilities of reidentification are still far too low to obtain any useful information.

8 Conclusions

Synthetic population data play an important part in the evaluation of statistical methods in the survey context. Without such data, it is not possible to perform design-based simulation studies.

This paper gives a brief outline of the flexible framework proposed in [5] for simulating population data for (household) surveys based on available sample data. The framework is applicable to a broad class of surveys and is implemented along with diagnostic plots in the R package `simPopulation`. In the case of

EU-SILC, the data generation procedure led to excellent results with respect to information loss.

In this paper, confidentiality issues of the generated synthetic EU-SILC population are discussed based on five different worst case scenarios. The results show that while reidentification is possible, an intruder would not gain any useful information from the purely synthetic data. Even if they successfully reidentify a unique combination of key variables from their data, the probability that the obtained value is close to the original value is extremely low for all considered worst case scenarios.

Due to our experiences and the results from the investigated scenarios, we can argue that synthetic population data generated with the methodology introduced in [5] and implemented in `simPopulation` are confidential and can be distributed to the public. Researchers could then use this data to evaluate the effects of different sampling designs, missing data mechanisms and outlier models on the estimator of interest in design-based simulation studies.

References

1. R. Münnich, J. Schürle, W. Bihler, H.-J. Boonstra, P. Knotterus, N. Nieuwenbroek, A. Haslinger, S. Laaksonen, D. Eckmair, A. Quatember, H. Wagner, J.-P. Renfer, U. Oetliker, and R. Wiegert. Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen (2003)
2. R. Chambers. Evaluation criteria for statistical editing and imputation. EurEdit Deliverable D3.3, Department of Social Statistics, University of Southhampton (2001)
3. A. Alfons, M. Templ, P. Filzmoser, S. Kraft, and B. Hulliger. Intermediate report on the data generation mechanism and on the design of the simulation study. AMELI Deliverable 6.1, Vienna University of Technology (2009)
4. R. Münnich and J. Schürle. On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen (2003)
5. A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of synthetic population data for household surveys with application to EU-SILC. Research Report CS-2010-1, Department of Statistics and Probability Theory, Vienna University of Technology (2010). `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf`
6. G.P. Clarke. Microsimulation: an introduction. In G.P. Clarke (ed.) *Microsimulation for Urban and Regional Policy Analysis*. Pion, London (1996)
7. D.B. Rubin. Discussion: Statistical disclosure limitation. *J. Off. Stat.*, 9(2):461–468 (1993)
8. R.J.A. Little. Statistical analysis of masked data. *J. Off. Stat.*, 9(2):407–426 (1993)
9. T.E. Raghunathan, J.P. Reiter, and D.B. Rubin. Multiple imputation for statistical disclosure limitation. *J. Off. Stat.*, 19(1):1–16 (2003)
10. J. Drechsler, S. Bender, and S. Rässler. Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Trans. Data Priv.*, 1(3):105–130 (2008)
11. J.P. Reiter. Using multiple imputation to integrate and disseminate confidential microdata. *Int. Stat. Rev.*, 77(2):179–195 (2009)

12. D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47(260):663–685 (1952)
13. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna (2010). ISBN 3-900051-07-0. `http://www.r-project.org`
14. S. Kraft and A. Alfons. *simPopulation: Simulation of synthetic populations for surveys based on sample data* (2010). R package version 0.1.2.
15. T. Atkinson, B. Cantillon, E. Marlier, and B. Nolan. *Social Indicators: The EU and Social Inclusion.* Oxford University Press, New York (2002). ISBN 0-19-925349-8.
16. Eurostat. Description of target variables: Cross-sectional and longitudinal. EU-SILC 065/04, Eurostat, Luxembourg (2004)
17. S. Kraft. Simulation of a Population for the European Living and Income Conditions Survey. Master's thesis, Vienna University of Technology (2009)
18. Y. Rinott and N. Shlomo. A generalized negative binomial smoothing model for sample disclosure risk estimation. In J. Domingo-Ferrer and L. Franconi (eds.) *Privacy in Statistical Databases*, *LNCS*, vol. 4302, pp. 82–93. Springer, Heidelberg (2006). ISBN 978-3-540-49330-3.
19. E.A.H. Elamir and C.J. Skinner. Record level measures of disclosure risk for survey microdata. *J. Off. Stat.*, 22(3):525–539 (2006)
20. M. Templ. *New Developments in Statistical Disclosure Control and Imputation: Robust Statistics Applied to Official Statistics.* Südwestdeutscher Verlag für Hochschulschriften, Germany (2009). ISBN 3838108280.
21. P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI International (1998)
22. L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Syst.*, 10(5):557–570 (2002)
23. L. Franconi and S. Polettini. Individual risk estimation in μ-ARGUS: A review. In J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases*, *LNCS*, vol. 3050, pp. 262–272. Springer, Heidelberg (2004). ISBN 978-3-540-22118-2.
24. J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.*, 14(1):189–201 (2002)
25. M. Templ and B. Meindl. Robust statistics meets SDC: New disclosure risk measures for continuous microdata masking. In J. Domingo-Ferrer and Y. Saygin (eds.) *Privacy in Statistical Databases*, *LNCS*, vol. 5262, pp. 113–126. Springer, Heidelberg (2008). ISBN 978-3-540-87470-6.
26. J.P. Reiter and R. Mitra. Estimating risks of identification disclosure in partially synthetic data. *J. Priv. Confid.*, 1(1):99–110 (2009)
27. G.T. Duncan and D. Lambert. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.*, 81(393):10–28 (1986)
28. S.E. Fienberg, U.E. Makov, and A.P. Sanil. A bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Off. Stat.*, 13(1):75–89 (1997)
29. H. Akaike. Statistical predictor identification. *Ann. Inst. Stat. Math.*, 22(2):203–217 (1970)