Stat Methods Appl manuscript No.
(will be inserted by the editor)

Simulation of close-to-reality population data for household surveys with application to EU-SILC

Andreas Alfons · Stefan Kraft · Matthias Templ · Peter Filzmoser

Received: date / Accepted: date

Abstract Statistical simulation in survey statistics is usually based on repeatedly drawing samples from population data. Furthermore, population data may be used in courses on survey statistics to explain issues regarding, e.g., sampling designs. Since the availability of real population data is in general very limited, it is necessary to generate synthetic data for such applications. The simulated data need to be as realistic as possible, while at the same time ensuring data confidentiality. This paper proposes a method for generating close-to-reality population data for complex household surveys. The procedure consists of four steps for setting up the household structure, simulating categorical variables, simulating continuous variables and splitting continuous variables into different components. It is not required to perform all four steps so that the framework is applicable to a broad class of surveys. In addition, the proposed method is evaluated in an application to the European Union Statistics on Income and Living Conditions (EU-SILC).

Keywords Synthetic data · Simulation · Survey statistics · EU-SILC

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

A. Alfons · S. Kraft · M. Templ · P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology
Wiedner Hauptstraße 7, 1040 Vienna, Austria
Tel.: +43 1 58801 10772
Fax: +43 1 58801 10798
E-mail: alfons@statistik.tuwien.ac.at

S. Kraft
now at the Institute for Quantitative Asset Management

M. Templ
Methods Unit, Statistics Austria

1 Introduction

Survey data contain variability due to sampling, imputation of missing values, measurement errors and editing. Statistical simulation in survey statistics therefore often follows a *close-to-reality* approach (see, e.g., Münnich et al 2003), i.e., the behavior of the developed methodology for a specific survey is investigated by repeatedly drawing samples from population data with the sampling method and weighting scheme used in practice. Population data may thus form the basis for a realistic framework to compare statistical methods under different settings. In particular, the estimation of indicators needed for policy decisions may be investigated with respect to different sampling designs or common data problems such as measurement errors or missing values.

In teaching, population data may support courses on topics such as sampling, statistical modeling or indicator estimation. Again, real-world situations could be considered by drawing samples from close-to-reality populations. Issues regarding, e.g., the sampling design or inhomogeneities in the data can be explained using real-world applications.

However, real population data are typically limited to census or register data. Only in exceptions are suitable population data available to researchers. The remedy of this problem is to generate synthetic populations from existing survey data.

Simulation of population microdata is closely related to the field of *microsimulation* (e.g., Clarke 1996), which is a well-established methodology within the social sciences, although the aims are quite different. Microsimulation models attempt to reproduce the behavior of individual units such as persons, households or firms over the course of many years for policy analysis purposes. Hence they are highly complex and time-consuming. Survey statisticians, on the other hand, need synthetic populations as a basis for extensive simulation studies on the behavior of their statistical methods. Fast computation is thus favored to over-complex models.

An alternative approach for the generation of synthetic data sets is discussed by Rubin (1993). He addresses the confidentiality problem connected with the release of publicly available microdata and proposes the generation of fully synthetic microdata sets using multiple imputation. Raghunathan et al (2003), Drechsler et al (2008) and Reiter (2009) discuss this approach in more detail. However, their approach does not allow to generate categories that are not represented in the original sample, nor do they investigate the possible generation of structural zeros in combinations of variables. Moreover, some basic variables from the real population data are required as auxiliary information.

The generation of population microdata for selected surveys as a basis for Monte Carlo simulation is described by Münnich et al (2003) and Münnich and Schürle (2003). Nevertheless, their framework was developed for household surveys with large sample sizes that contain mainly categorical variables. All steps of the procedure are performed separately for each stratum of the sampling design. The household structure is thereby simulated in two steps. First, the household sizes are drawn from the observed conditional distribu-

tions within the strata. Second, the age and gender structure of the population households is generated by resampling households of the same size from the respective strata in the sample. Additional categorical variables are then simulated by random draws from the observed conditional distributions of their multivariate realizations within each combination of stratum, age (or age category) and gender. Also continuous variables are modeled separately for each combination of stratum and outcomes from certain influential variables.

In any case, this framework has been modified and extended in order to be applicable to more complex surveys such as the well-known *European Union Statistics on Income and Living Conditions* (EU-SILC). Please note that while it would be interesting to establish a theoretical relationship between the goodness of the statistical models and the resulting populations, such an analysis is out of scope for this paper due to the large number of models involved. Instead, the proposed procedure is evaluated by means of simulation.

The rest of the paper is organized as follows. In Section 2, the proposed data simulation method is described in great detail. Diagnostic plots and results from extensive simulation studies in an application to EU-SILC are presented in Section 3. The final Section 4 concludes.

2 Simulation of synthetic populations

The data simulation method proposed in this paper is motivated by the *European Union Statistics on Income and Living Conditions* (EU-SILC; see Section 3), but since it is designed to manage all difficulties of this highly complex survey, it is also applicable to many other household surveys. In any case, the following conditions need to be respected when simulating population data (Münnich et al 2003; Münnich and Schürle 2003):

- Actual sizes of regions and strata need to be reflected.
- Marginal distributions and interactions between variables should be represented correctly.
- Heterogeneities between subgroups, especially regional aspects, should be allowed.
- Pure replication of units from the underlying sample should be avoided, as this generally leads to extremely small variability of units within smaller subgroups.
- Data confidentiality must be ensured.

In the case of EU-SILC, another problem needs to be considered. As the name suggests, EU-SILC contains information about income, which is split into different income components. The data simulation method must thus ensure that such a breakdown of variables is done in a realistic manner.

Since some of the above conditions are conflicting with one another, generating completely realistic populations seems an impossible task. Nevertheless, being as close to reality as possible suffices for drawing meaningful conclusions from simulation studies.

Our procedure is based on the ideas of Münnich et al (2003) and Münnich and Schürle (2003). However, they mainly consider the generation of categorical variables for specific surveys such as the German Microcensus, with only a few simple extensions to continuous variables. The proposed method uses modifications of their framework and both improves and extends the simulation scheme such that it can be applied to a much broader class of household surveys. This in particular includes surveys with relatively small sample sizes or with complex continuous variables or components thereof. In general, the procedure consists of four steps:

1. Setup of the household structure
2. Simulation of categorical variables
3. Simulation of continuous variables
4. Splitting continuous variables into components

While the propositions of Münnich et al (2003) and Münnich and Schürle (2003) are only slightly modified in Step 1, an entirely different approach is used in Steps 2 and 3. In addition, Step 4 constitutes a new development motivated by EU-SILC. Having different stages provides maximum flexibility of the framework. Depending on the specific survey, not all four steps need to be carried out.

It is important to note that the proposed data generation method relies solely on the underlying sample data, no auxiliary information (e.g., available census data) is required. Stratification allows to account for heterogeneities such as regional differences. Furthermore, sample weights are considered in each step to ensure high similarity of expected and realized values. Concerning data confidentiality, a detailed analysis of the framework using different worst case scenarios is carried out in Templ and Alfons (2010). The conclusion of this analysis is that the synthetic population data are confidential and may be distributed to the public.

In the following sections, the different steps of the procedure are described in detail. Section 2.5 then briefly discusses the implementation of the procedure in R (R Development Core Team 2010).

2.1 Setup of the household structure

The household structure is simulated separately for each combination of stratum k and household size l. First, the number of households M_{kl} is estimated using the Horvitz-Thompson estimator (Horvitz and Thompson 1952):

$$\hat{M}_{kl} := \sum_{h \in H_{kl}^S} w_h, \tag{1}$$

where H_{kl}^S denotes the index set of households in stratum k of the survey data with household size l, and w_h, $h \in H_{kl}^S$, are the corresponding household weights. Similarly, let H_{kl}^U be the respective index set of households in the

population data such that $|H_{kl}^U| = \hat{M}_{kl}$. To prevent unrealistic structures in the population households, basic information from the survey households is resampled. Let x_{hij}^S and x_{hij}^U denote the value of person i from household h in variable j for the sample and population data, respectively, and let the first p_1 variables contain the basic information on the household structure. For each population household $h \in H_{kl}^U$, a survey household $h' \in H_{kl}^S$ is selected with probability $w_{h'}/\hat{M}_{kl}$ and the household structure is set to

$$x_{hij}^U := x_{h'ij}^S, \quad i = 1, \ldots, l, \; j = 1, \ldots, p_1. \tag{2}$$

Alias sampling (Walker 1977) is well suited for our purpose, as it is very fast for a large number of sampled elements. Furthermore, as few variables as possible should be adopted by the persons in the resampled households for disclosure reasons. Our suggestion is to use only age and gender information, which is typically available in household surveys.

2.2 Simulation of categorical variables

For simulating additional categorical variables, the approach by Münnich et al (2003) and Münnich and Schürle (2003) is based on estimating conditional distributions directly by the corresponding relative frequency distributions in the underlying sample. It therefore requires a rather large sample size and is not very flexible (see Section 3.2; cf. Kraft 2009). In particular, it does not allow to generate combinations that do not occur in the sample. To overcome these shortcomings, the proposed approach estimates conditional distributions with multinomial logistic regression models.

Let $\boldsymbol{x}_j^S = (x_{1j}^S, \ldots, x_{nj}^S)'$ and $\boldsymbol{x}_j^U = (x_{1j}^U, \ldots, x_{Nj}^U)'$ denote the variables in the sample and population, respectively, where n and N give the corresponding number of individuals. The additional categorical variables are thereby given by the indices $p_1 < j \leq p_2$. Furthermore, the personal sample weights are denoted by $\boldsymbol{w} = (w_1, \ldots, w_n)'$. Multinomial logistic regression models are fitted for each stratum separately. Due to limited space, a detailed mathematical description of these models cannot be provided in this paper, but can be found in, e.g., Simonoff (2003).

The following procedure is performed for each stratum k and each variable to be simulated, given by the index j, $p_1 < j \leq p_2$. Let I_k^S and I_k^U be the index sets of individuals in stratum k for the survey and population data, respectively. The survey data given by the indices in I_k^S is used to fit the model with response $\boldsymbol{x}_j^S$ and predictors $\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_{j-1}^S$, thereby considering the sample weights w_i, $i \in I_k^S$. Furthermore, let $\{1, \ldots, R\}$ be the set of possible outcome categories of the response variable. In particular, the number of possible outcomes is denoted by R. For every individual $i \in I_k^U$, the conditional

probabilities $p_{ir}^U := P(x_{ij}^U = r | x_{i1}^U, \ldots, x_{i,j-1}^U)$ are estimated by

$$\begin{aligned} \hat{p}_{i1}^U &:= \frac{1}{1 + \sum_{r=2}^R \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}, \\ \hat{p}_{ir}^U &:= \frac{\exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}{1 + \sum_{r=2}^R \exp(\hat{\beta}_{0r} + \hat{\beta}_{1r} x_{i1}^U + \ldots + \hat{\beta}_{j-1,r} x_{i,j-1}^U)}, \quad r = 2, \ldots, R, \end{aligned} \tag{3}$$

where $\hat{\beta}_{0r}, \ldots, \hat{\beta}_{j-1,r}$, $r = 2, \ldots, R$, are the estimated coefficients (see, e.g., Simonoff 2003). The values of $\boldsymbol{x}_j^U$ for the individuals $i \in I_k^U$ are then drawn from the corresponding conditional distributions.

Note that for simulating the jth variable, $p_1 < j \leq p_2$, the $j-1$ previous variables are used as predictors. This means that the order of the additional categorical variables may be relevant. However, once such a variable is generated in the population, that information should certainly be used for simulating the remaining variables. In our application to EU-SILC, changing the order of the variables did not produce significantly different results (not shown). Alternatively, the procedure could be continued iteratively once all additional variables are available in the population, in each step using all other variables as predictors. Nevertheless, such a procedure would be computationally very expensive for real-life sized population data.

Estimating the conditional distributions with multinomial logistic regression models allows to simulate combinations that do not occur in the sample but are likely to occur in the true population. Such combinations are called *random zeros*, as opposed to *structural zeros*, which are impossible to occur (e.g., Simonoff 2003). For close-to-reality populations, such structural zeros need to be reflected. This can be done by setting $p_{ir'}^U := 0$, where r' is an impossible value for x_{ij} given $x_{i1}, \ldots, x_{i,j-1}$, and adjusting the other probabilities so that $\sum_{r=1}^R p_{ir}^U = 1$.

Keep in mind that the idea of the proposed data simulation framework is to proceed in a stepwise fashion, generating different types of variables in each step. However, the procedure could easily be modified to allow for previously simulated continuous predictors when simulating a categorical variable.

2.3 Simulation of continuous variables

Continuing the notation from the previous section, let $\boldsymbol{x}_j^S$ and $\boldsymbol{x}_j^U$, $p_2 < j \leq p_3$, denote the continuous variables. Two different approaches are presented in the following. Both are able to handle semi-continuous variables, i.e., variables that contain a large amount of zeros.

2.3.1 Multinomial model with random draws from resulting categories

This approach is based on the simulation of categorical variables described in the previous section. The following steps are performed for each variable to

be simulated, given by the index j, $p_2 < j \leq p_3$. First, the variable $\boldsymbol{x}_j^S$ is discretized. This is done in a different manner for continuous and semi-continuous variables. For continuous variables, $R+1$ breakpoints $b_1 < \ldots < b_{R+1}$ are used to define the discretized variable $\boldsymbol{y}^S = (y_1^S, \ldots, y_n^S)'$ as

$$y_i^S := \begin{cases} 1 & \text{if } b_1 \leq x_{ij}^S \leq b_2, \\ r & \text{if } b_r < x_{ij}^S \leq b_{r+1},\ r = 2, \ldots, R. \end{cases} \tag{4}$$

For semi-continuous variables, zero is a category of its own, and breakpoints for negative and positive values are distinguished. Let $b_{R^-ยน+1}^- < \ldots < b_1^- = 0 = b_1^+ < \ldots < b_{R^+ยน+1}^-$ be the breakpoints. Then $\boldsymbol{y}^S$ is defined as

$$y_i^S := \begin{cases} -r & \text{if } R^- > 0 \text{ and } b_{r+1}^- \leq x_{ij}^S < b_r^-,\ r = R^-, \ldots, 1, \\ 0 & \text{if } x_{ij}^S = 0, \\ r & \text{if } R^+ > 0 \text{ and } b_r^+ < x_{ij}^S \leq b_{r+1}^+,\ r = 1, \ldots, R^+. \end{cases} \tag{5}$$

Note that the cases of only non-negative or non-positive values in $\boldsymbol{x}_j^S$ are considered in (5).

Multinomial logistic regression models with response $\boldsymbol{y}^S$ and predictors $\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_{j-1}^S$ are then fitted for every stratum k separately, as described in the previous section, in order to simulate the values of the categorized population variable $\boldsymbol{y}^U = (y_1^U, \ldots, y_N^U)'$.

Finally, the values of $\boldsymbol{x}_j^U$ are generated by random draws from uniform distributions within the corresponding categories of $\boldsymbol{y}^U$. For continuous variables, the values of individual $i = 1, \ldots, N$ are generated as

$$x_{ij}^U \sim U(b_r, b_{r+1}) \text{ if } y_i^U = r. \tag{6}$$

For semi-continuous variables, the values of individual $i = 1, \ldots, N$ are set to $x_{ij}^U := 0$ if $y_i^U = 0$, while the non-zero observations are generated as

$$x_{ij}^U \sim \begin{cases} U(b_{r+1}^-, b_r^-) & \text{if } y_i^U = -r < 0, \\ U(b_r^+, b_{r+1}^+) & \text{if } y_i^U = r > 0. \end{cases} \tag{7}$$

The idea behind this approach is to divide the data into relatively small subsets. If the intervals are too large, using uniform distributions may be an oversimplification. However, the advantage of this approach is that it allows the breakpoints for the discretization to be chosen in such a way that the empirical distribution is well reflected in the simulated population variable. It thereby needs to be considered that the larger the number of breakpoints, the higher the computation time. Quantiles in steps of 10% are reasonable default values for the breakpoints, while the fit in the tails of the distribution may be improved by also using the 1%, 5%, 95% and 99% quantiles. Note that sufficient accuracy in some applications may already be reached with larger steps in the middle part of the distribution (see Section 3).

When simulating variables that contain extreme values, such as income, *tail modeling* should be considered. In that case, values from the largest categories

could be drawn from a generalized Pareto distribution (GPD). The cumulative distribution function of the GPD is defined as

$$F_{\mu,\sigma,\xi}(x) = \begin{cases} 1 - \left(1 + \dfrac{\xi(x-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi = 0, \end{cases}$$

where μ is the location parameter, $\sigma > 0$ is the scale parameter and ξ is the shape parameter. The range of x is $x \geq 0$ when $\xi \geq 0$ and $\mu \leq x \leq \mu - \frac{\sigma}{\xi}$ when $\xi < 0$. See, e.g., Embrechts et al (1997) for details on the *peaks over threshold* approach for fitting the GPD. Note that other distributions may be used for tail modeling as well (see, e.g., Kleiber and Kotz 2003). Nevertheless, if the purpose of such a population is comparing different estimators in a simulation study, it is important to note that using a GPD for the tails favors estimators that incorporate generalized Pareto tail modeling over other types of estimators.

2.3.2 (Two-step) regression model with random error terms

The second approach is based on linear regression combined with random error terms. Semi-continuous variables are thereby simulated using a two-step model. The following procedure is repeated for each variable to be simulated, given by the index j, $p_2 < j \leq p_3$.

For semi-continuous variables, the first step is to simulate whether x_{ij}^U, $i = 1, \ldots, N$, is zero or not. This is done by fitting logistic regression models (see, e.g., Simonoff 2003) for each stratum separately. The binary response variable $\boldsymbol{y}^S = (y_1^S, \ldots, y_n^S)'$ is defined as

$$y_i^S := \begin{cases} 0 & \text{if } x_{ij} = 0, \\ 1 & \text{else.} \end{cases} \tag{8}$$

For each stratum k, the observations given by the index set I_k^S are used to fit the model with response $\boldsymbol{y}^S$ and predictors $\boldsymbol{x}_1^S, \ldots, \boldsymbol{x}_{j-1}^S$. The sample weights w_i, $i \in I_k^S$, are considered in the model fitting process by using a weighted maximum likelihood approach. For every individual $i \in I_k^U$, the conditional probabilities $p_i^U := P(y_i^U = 1 | x_{i1}^U, \ldots, x_{i,j-1}^U)$ that x_{ij}^U is non-zero are estimated by

$$\hat{p}_i^U := \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}^U + \ldots + \hat{\beta}_{j-1} x_{i,j-1}^U)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}^U + \ldots + \hat{\beta}_{j-1} x_{i,j-1}^U)}, \tag{9}$$

where $\hat{\beta}_0, \ldots, \hat{\beta}_{j-1}$ are the estimated coefficients (e.g., Simonoff 2003). The values y_i^U, $i \in I_k^U$, are then drawn from the corresponding conditional distributions. Consequently, the zeros in the simulated semi-continuous variable are given by $x_{ij}^U := 0$ if $y_i^U = 0$. For the second step, the non-zero observations are indicated by $\tilde{I}_k^S := \{i \in I_k^S : y_i^S = 1\}$ and $\tilde{I}_k^U := \{i \in I_k^U : y_i^U = 1\}$.

For continuous variables, on the other hand, $\tilde{I}_k^S := I_k^S$ and $\tilde{I}_k^U := I_k^U$ are used in the following. Linear regression models are fitted for every stratum separately. In order to obtain more robust models, trimming parameters α_1 and α_2 are introduced. The following procedure is carried out for each stratum k. Let the observations to be used for fitting the model be given by the index set $I_{\alpha_1}^{\alpha_2} := \{i \in \tilde{I}_k^S : q_{\alpha_1} < x_{ij} < q_{1-\alpha_2}\}$, where q_{α_1} and $q_{1-\alpha_2}$ are the corresponding α_1 and $1-\alpha_2$ quantiles, respectively. The linear model is then given by

$$x_{ij}^S = \beta_0 + \beta_1 x_{i1}^S + \ldots + \beta_{j-1} x_{i,j-1}^S + \varepsilon_i^S, \quad i \in I_{\alpha_1}^{\alpha_2}, \tag{10}$$

where ε_i^S are random error terms. Using the weighted least squares approach with weights w_i, $i \in I_{\alpha_1}^{\alpha_2}$, coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_{j-1}$ are obtained (see, e.g., Weisberg 2005) and the population values are estimated by

$$\hat{x}_{ij}^U = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}^U + \ldots + \hat{\beta}_{j-1} x_{i,j-1}^U + \varepsilon_i^U, \quad i \in \tilde{I}_k^U. \tag{11}$$

The random error terms ε_i^U need to be added since otherwise individuals with the same set of predictor values would receive the same value in x_{ij}^U. There are two suggestions on how to generate the random error terms:

- Use random draws from the residuals

$$\hat{r}_i^S = x_{ij}^S - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}^S + \ldots + \hat{\beta}_{j-1} x_{i,j-1}^S\right), \quad i \in I_{\alpha_1}^{\alpha_2}. \tag{12}$$

- Use random draws from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. The parameters μ and σ are thereby estimated robustly with median and MAD, respectively.

The first approach is more data-driven, while the second approach is in accordance with the theoretical assumption of normally distributed errors. For both, the trimming parameters α_1 and α_2 need to be selected carefully. If they are too small, very large random error terms due to outliers may result in large deviations especially in the tails of the distribution. If they are too large, the random error terms may not introduce enough variability. In the application to EU-SILC, $\alpha_1 = \alpha_2 = 0.01$ appeared to be a reasonable choice.

For variables such as income, a log-transformation may be beneficial before fitting the linear model. Equation (10) is then changed to

$$\log x_{ij}^S = \beta_0 + \beta_1 x_{i1}^S + \ldots + \beta_{j-1} x_{i,j-1}^S + \varepsilon_i^S, \quad i \in I_{\alpha_1}^{\alpha_2}. \tag{13}$$

In that case, the population values are estimated by

$$\hat{x}_{ij}^U = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}^U + \ldots + \hat{\beta}_{j-1} x_{i,j-1}^U + \varepsilon_i^U), \quad i \in \tilde{I}_k^U. \tag{14}$$

However, the log-transformation causes problems with negative values, which is realistic for income (losses from self employment, see the example with EU-SILC data in Section 3). A simple remedy is of course to add a constant $c > 0$ to x_{ij}^S to obtain positive values, i.e., to use $\log(x_{ij}^S + c)$ in the left-hand side of Equation (13). This constant then needs to be subtracted from the right hand side of Equation (14). Another possibility is to combine the two presented

approaches for simulating (semi-)continuous variables. A multinomial model with one category for positive values and certain categories for non-positive values is applied in the first step. Positive values are then simulated using a linear model for the log-transformed data, while negative values are drawn from uniform distributions within the respective simulated categories.

2.4 Splitting continuous variables into components

The procedure for simulating components of continuous variables is motivated by EU-SILC data, which contain information on various income components. When simulating components, the following problems need to be considered (cf. Kraft 2009). Even for a moderate number of components, it may be too complex to consider all the dependencies between the components and the other variables, as well as between the components themselves. Moreover, sparseness of various components may be an issue, e.g., in EU-SILC data, most income components typically contain only few non-zero observations. To manage these problems, a simple but effective approach based on conditional resampling of fractions has been developed. Only very few highly influential categorical variables should thereby be considered for conditioning.

Let $\boldsymbol{z}^S = (z_1^S, \ldots, z_n^S)'$ and $\boldsymbol{z}^U = (z_1^U, \ldots, z_N^U)'$ denote the variable giving the total in the sample and population, respectively, and let $\boldsymbol{x}_j^S$ and $\boldsymbol{x}_j^U$, $p_3 < j \leq p_4$, denote the variables containing the components. First, the fractions of the components with respect to the total are computed for the sample:

$$y_{ij}^S := \frac{x_{i,p_3+j}^S}{z_i^S}, \quad i \in I_r^S,\ j = 1, \ldots, p_4 - p_3. \tag{15}$$

For the second step, let J_c be the index set of the conditioning variables. This step is performed separately for every combination of outcomes $\boldsymbol{r} = (r_j)_{j \in J_c}$. Let $I_r^S := \{i : x_{ij}^S = r_j\ \forall j \in J_c\}$ and $I_r^U := \{i : x_{ij}^U = r_j\ \forall j \in J_c\}$ be the index sets of individuals in the survey and population data, respectively, with the corresponding outcomes in the conditioning variables. For each individual $i \in I_r^U$ in the population, an individual $i' \in I_r^S$ from the survey data is selected with probability $w_{i'} / \sum_{i \in I_r^S} w_i$ and the values of the components are set to

$$x_{i,p_3+j}^U := z_i^U y_{i'j}^S, \quad j = 1, \ldots, p_4 - p_3. \tag{16}$$

If no observations for combination $\boldsymbol{r}$ exist in the sample, i.e., if $I_r^S = \emptyset$, a suitable donor $\boldsymbol{r}'$ is selected by minimizing a suitable distance measure such as the Manhattan distance $d_1(\boldsymbol{r}, \boldsymbol{s}) = \|\boldsymbol{r} - \boldsymbol{s}\|_1$. Then $I_r^S := I_{r'}^S$ is used in the above steps.

Resampling fractions has the advantage that it avoids unrealistic or unreasonable combinations in the simulated components. At the same time, it does not result in pure replication, as the absolute values for simulated individuals are in general quite different from the corresponding individuals in the underlying survey data.

2.5 Software

The proposed data simulation framework is implemented in the R package `simPopulation` (Alfons and Kraft 2010), which can be obtained from CRAN (the Comprehensive R Archive Network, `http://cran.r-project.org`). For maximum flexibility, the four steps of the procedure are available as separate functions. To generate populations for EU-SILC, a wrapper combining all four steps is implemented in order to provide a more convenient interface. Wrappers for other surveys can easily be defined by the user. In addition, functions to create diagnostic plots as shown in Section 3.1 are available. The latter are implemented using packages `vcd` (Meyer et al 2006, 2010) and `lattice` (Sarkar 2008, 2011).

It would certainly be beneficial to present a line-by-line illustration of the R code for the application in Section 3. Nevertheless, the EU-SILC sample provided by Statistics Austria is confidential, thus the reader would not be able to reproduce the results. Furthermore, the additional explanation of the R code would render the length of the paper far from being reasonable. Therefore, detailed instructions for such an analysis and the generation of diagnostic plots are provided in a separate package vignette (Alfons et al 2010b). If `simPopulation` is installed, the vignette can be viewed from within R with the following command:

```
R> vignette("simPopulation-eusilc")
```

Note that the vignette uses the synthetically generated example data from the package, hence the results presented there are reproducible.

3 Application to EU-SILC

The *European Union Statistics on Income and Living Conditions* (EU-SILC) is one of the most well-known panel surveys and is conducted in EU member states and other European countries. It is mainly used as data basis for the *Laeken indicators*, a set of indicators for measuring risk-of-poverty and social cohesion in European countries (cf. Atkinson et al 2002).

The application of the proposed data simulation procedure to EU-SILC (limited to non-negative personal net income and income components) is described in more detail in Kraft (2009), where an extensive collection of results can be found as well. With the generalizations presented in this paper, however, it is also possible to simulate negative income or income components. The underlying survey data used in this section is the Austrian EU-SILC sample from 2006. Table 1 lists the variables to be included in the simulation and their possible outcomes. It should be noted that due to low frequencies of occurence, some categories of economic status and citizenship, respectively, have been combined. Such combined categories are marked with an asterisk (*) in Table 1. A complete description of variables in EU-SILC and possible outcomes can be found in Eurostat (2004).

Table 1 Variables selected for the simulation of the Austrian EU-SILC population data.

Variable	**Name**	**Possible outcomes**	
Region	`db040`	1	Burgenland
		2	Lower Austria
		3	Vienna
		4	Carinthia
		5	Styria
		6	Upper Austria
		7	Salzburg
		8	Tyrol
		9	Vorarlberg
Household size	`hsize`		Number of persons in household
Age	`age`		Age (for the previous year) in years
Gender	`rb090`	1	Male
		2	Female
Self-defined current economic status	`pl030`	1	Working full-time
		2	Working part-time
		3	Unemployed
		4	Pupil, student, further training or unpaid work experience or in compulsory military or community service*
		5	In retirement or in early retirement or has given up business
		6	Permanently disabled or/and unfit to work or other inactive person*
		7	Fulfilling domestic tasks and care responsibilities
Citizenship	`pb220a`	1	Austria
		2	EU*
		3	Other*
Personal net income	`netIncome`		Sum of income components listed below
Employee cash or near cash income	`py010n`	0	No income
		> 0	Income
Cash benefits or losses from self-employment	`py050n`	< 0	Losses
		0	No income
		> 0	Benefits
Unemployment benefits	`py090n`	0	No income
		> 0	Income
Old-age benefits	`py100n`	0	No income
		> 0	Income
Survivor's benefits	`py110n`	0	No income
		> 0	Income
Sickness benefits	`py120n`	0	No income
		> 0	Income
Disability benefits	`py130n`	0	No income
		> 0	Income
Education-related allowances	`py140n`	0	No income
		> 0	Income

* combined categories

Section 3.1 presents some diagnostic plots for comparing synthetic population data to the underlying sample. How well the characteristics of the original sample are reflected in such synthetic populations is further assessed by simulation in Section 3.2. These comparisons with the underlying sample are essential as this is the only real data available. Weighted distributions are thereby used for the sample data in all comparisions. In Section 3.3, the influence of different sample sizes and sampling designs on the proposed methodology is investigated by more extensive simulation studies.

3.1 Diagnostic plots for a single simulation

For setting up the household structure, households from the survey data are resampled conditional on region and household size. Sensible correlation structures within the households are ensured by resampling the variables age and gender, as recommended in Section 2.1. Afterwards, the variable age is categorized in order to use the resulting categories for the rest of the simulation. Variables that are categorized in the data simulation procedure are listed in Table 2, along with the respective categories. Besides age, the personal net income is discretized at a later stage. The age categories are chosen as a resonable tradeoff between accuracy and computation time (see also Section 3.2). Children below 16 are combined into one category since EU-SILC provides information for the remaining variables to be simulated only for persons of age 16 or above (see Eurostat 2004). Furthermore, one category for all persons of age above 80 is used due to the low frequencies of occurrence. In any case, economic status and citizenship are simulated for every region separately. In the multinomial logistic regression models described in Section 2.2, the predictors age category, gender and household size are used for economic status, while age category, gender, household size and economic status are then used to simulate citizenship.

In this section, the structure of the simulated categorical variables is evaluated by graphical means only. Figure 1 contains mosaic plots visualizing the expected and realized frequencies of gender, region and household size (*top*), as well as gender, economic status and citizenship (*bottom*). Both show very

Table 2 Categorized variables created for use as predictors during the simulation.

Variable	**Name**	**Categories**
Age category	`ageCat`	≤ 15, $(15, 20]$, $(20, 25]$, $(25, 30]$, $(30, 35]$, $(35, 40]$, $(40, 45]$, $(45, 50]$, $(50, 55]$, $(55, 60]$, $(60, 65]$, $(65, 70]$, $(70, 75]$, $(75, 80]$, > 80
Personal net income category	`netIncomeCat`	$[-9600, -5840)$, $[-5840, -4200)$, $[-4200, 0)$, 0, $(0, 800]$, $(800, 2800]$, $(2800, 5021.56]$, $(5021.56, 8456]$, $(8456, 13720]$, $(13720, 17738]$, $(17738, 23601.65]$ $(23601.65, 29191.86]$, $(29191.86, 36000]$, $(36000, 57227.69]$, > 57227.69

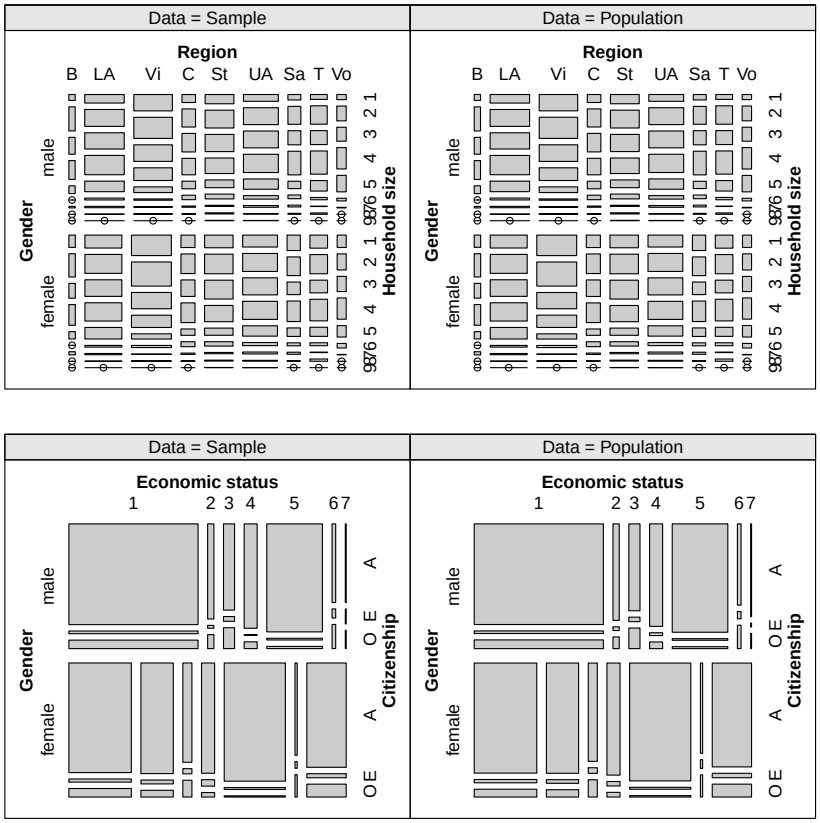

Fig. 1 *Top*: Mosaic plots of gender, region and household size. *Bottom*: Mosaic plots of gender, economic status and citizenship.

similar structures in the sample and population data. Note that these plots have been selected representatively, as the number of possible combinations of variables is too large to show them all. However, the interactions between all categorical variables are very well reflected in the synthetic population data. This is further documented in Section 3.2 by average relative differences of contingency coefficients from multiple simulation runs. While the two plots at the top of Figure 1 are nearly identical, closer inspection of the two plots at the bottom reveals small differences. These differences are due to the multinomial logistic regression models. The following two points need to be kept in mind. First, the expected frequencies of the different combinations are solely determined by the sum of the corresponding sample weights. Second, the multinomial models allow simulating combinations that do not occur in the sample but are likely to occur in the population. Consequently, the differences may be interpreted as corrections of the expected frequencies. For additional results

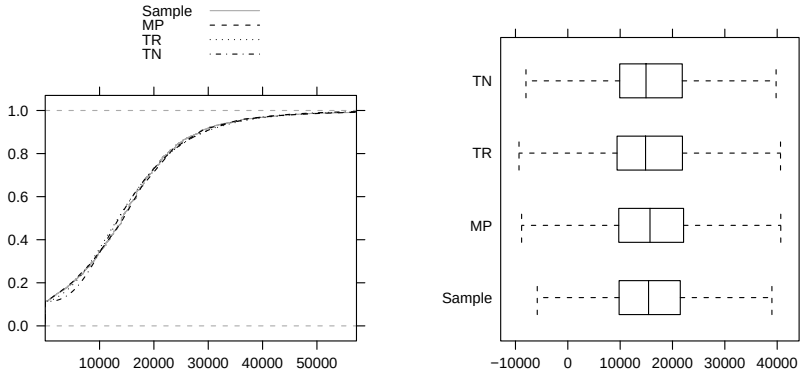

Fig. 2 *Left*: Cumulative distribution functions of personal net income. For better visibility, the plot shows only the main parts of the data. *Right:* Box plots of personal net income. Points outside the extremes of the whiskers are not plotted.

concerning the simulation of categorical variables in the case of EU-SILC, including χ^2 goodness of fit tests, the reader is referred to Kraft (2009).

For simulating personal net income, the two approaches described in Section 2.3 are compared. In both cases, the variables age category, gender, household size, economic status and citizenship are used as predictors and the models are computed separately for each region. The approach based on multinomial logistic regression models thereby uses the following parameter settings. In the categorization of personal net income, zero is a category of its own since personal net income is a semi-continuous variable. Breakpoints for the positive values are chosen as their weighted 1%, 5%, 10%, 20%, 40%, 60%, 80%, 90%, 95% and 99% quantiles. Furthermore, the only three negative values are used as breakpoints for negative income. See Table 2 for the resulting income categories. Values in the categories above the two largest breakpoints are drawn from a truncated generalized Pareto distribution. In the following, this approach will be referred to as *MP*. For the two-step linear regression approach, on the other hand, two different parameter settings are investigated. The first uses random draws from the residuals and will be referred to as *TR*, the second uses random draws from a normal distribution and will be referred to as *TN*. In both cases, the positive sample data are trimmed with parameters $\alpha_1 = \alpha_2 = 0.01$ and log-transformed in the second step of the procedure. Trimming is used since this performed better (results not shown, cf. Kraft 2009). In order to simulate negative income, a multinomial model is used in the first step. For negative income, again the only three existing values are used as breakpoints (see Table 2), and the simulated values are drawn from uniform distributions in the corresponding classes.

In Figure 2 (*left*), the cumulative distribution functions (CDF) of personal net income in the three simulated populations are compared to the empiri-

cal CDF obtained from the sample. Sample weights are taken into account by adjusting the step height. For better visibility of the differences, the plot shows only the main parts of the data (from 0 to the weighted 99% quantile of the positive values in the sample). The CDFs indicate an excellent fit, in particular for the MP approach. With the TN approach, there are some deviations for lower income, though. Figure 2 (*right*) uses box plots to compare the distributions. The box plots are adapted for semi-continuous variables in the following way. Box and whiskers are computed only for the non-zero part of the data and the box widths are proportional to the ratio of non-zero observations to the total number of observed values. For the sample data, sample weights are taken into account when computing the box plot statistics and the box widths. These box plots suggest that the proposed approaches perform well regarding the proportion of individuals with zero income and the distribution of non-zero income for the main part of the data.

Figure 3 contains box plots of the conditional distributions of personal net income with respect to gender (*top left*), citizenship (*top right*), region (*bottom left*) and economic status (*bottom right*). The proportions of zeros and the distributions of the non-zero observations appear to be in general well reflected in the simulated populations. Only some very small subgroups of economic status show significant deviations for the two-step regression approaches. This underlines the good fit of the models and illustrates that the proposed methods succeed to account for heterogeneities in the data.

Last but not least, the income components are simulated conditional on income category and economic status (see Section 2.4). Box plots of the results are shown in Figure 4. Due to the large number of zeros in most income components, a minimum box width is used in some cases to prevent the corresponding boxes from deteriorating into lines. In any case, the plots suggest that the simulation procedure for splitting variables into components works very well.

Additional results from simulations restricted to non-negative income, including correlation coefficients of the income components, can be found in Kraft (2009).

3.2 Average results from multiple simulations

In this section, the quality of the proposed methods is further assessed by simulation. With the parameter settings as described in the previous section, 100 populations are simulated. Certain quantities of interest for the sample data are thereby compared to the averages of their population counterparts over all simulation runs. The R package `simFrame` (Alfons et al 2010a; Alfons 2010) is used to manage the multiple simulations.

The relationships between the categorical variables, including the variables defining the household structure (age, gender and household size), are evaluated using contingency coefficients. Pearson's coefficient of contingency is a measure of association for categorical data defined as $P = \sqrt{\chi^2/(n + \chi^2)}$,

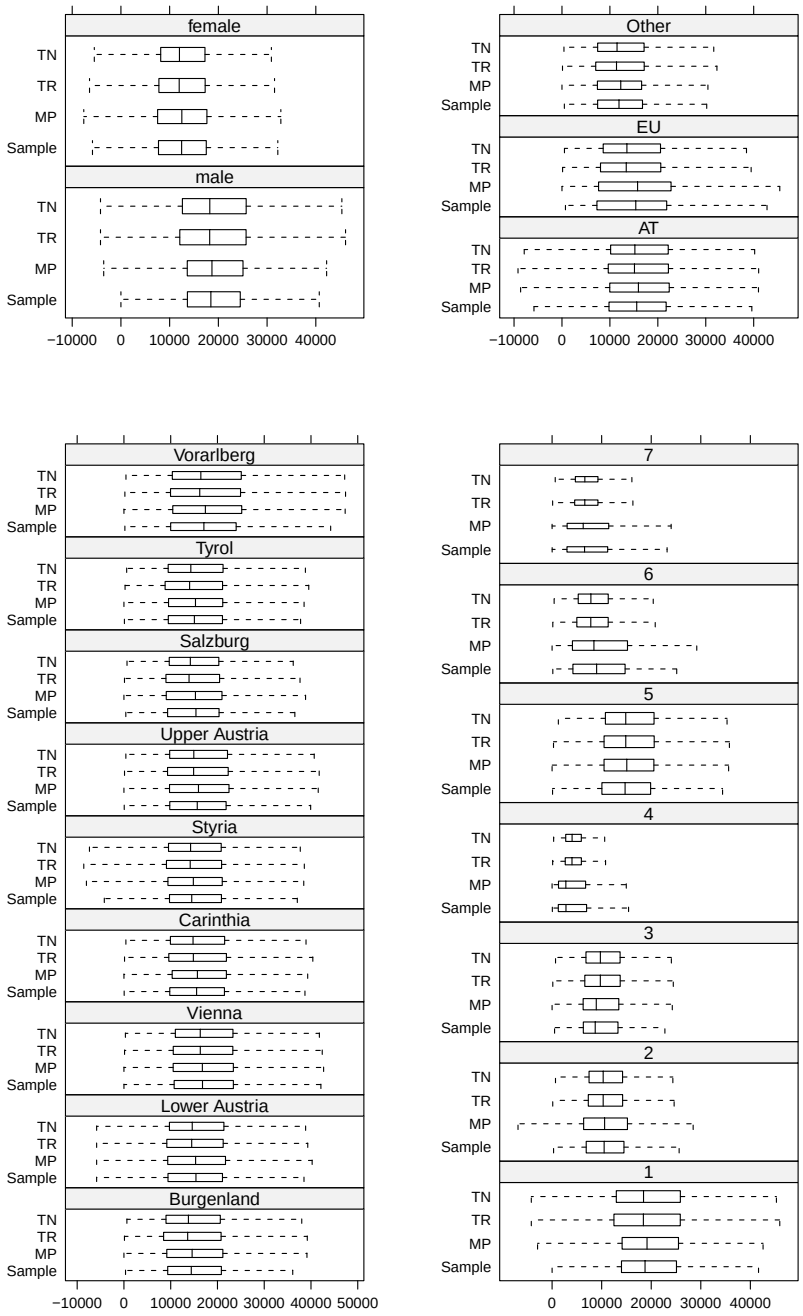

Fig. 3 Box plots of personal net income split by gender (*top left*), citizenship (*top right*), region (*bottom left*) and economic status (*bottom right*). Points outside the extremes of the whiskers are not plotted.

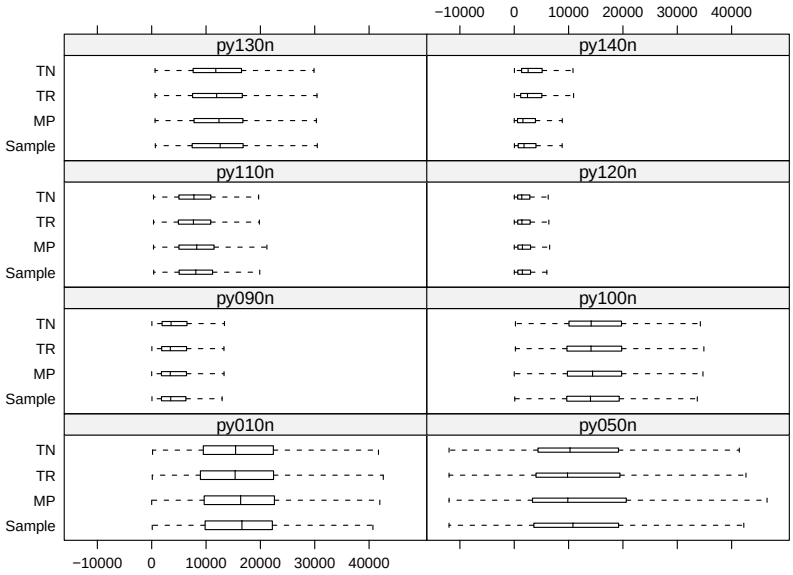

Fig. 4 Box plots of the income components. Points outside the extremes of the whiskers are not plotted.

where χ^2 is the test statistic of the χ^2 test of independence and n is the number of observations (see, e.g., Kendall and Stuart 1967, for more information).

Furthermore, the proposed methodology for the generation of categorical variables is compared to the framework of Münnich et al (2003) and Münnich and Schürle (2003). For the simulation of the household structure, the household sizes in our procedure are obtained in a completely deterministic way by estimated population totals, whereas they draw the household sizes from the observed conditional distributions within the strata. However, the age and gender structure is also generated by resampling households from the corresponding strata. Additional categorical variables are in their framework then simulated by random draws from the observed conditional distributions of the multivariate realizations within each combination of stratum, age category and gender. Keep in mind that this does not allow to simulate combinations that do not occur in the sample.

Table 3 compares the contingency coefficients obtained from the sample to the average results over the simulation runs. Note that the values for the sample are based on weighted distributions. For the proposed procedure, the relative differences are negligible, except for the coefficient of age and citizenship (`pb220a`). This exception is a result of using age categories for the prediction of citizenship, which can be avoided by using more categories or the original uncategorized age information. On the other hand, this increases

Table 3 Pairwise contingency coefficients of the categorical variables for the sample data (*top*), as well as average results from 100 simulated populations using the proposed method (*middle*), and the method of Münnich et al (2003) and Münnich and Schürle (2003) (*bottom*).

		age	rb090	hsize	pl030	pb220a
Sample	db040	0.261	0.019	0.217	0.139	0.161
	age		0.118	0.546	0.723	0.194
	rb090			0.081	0.385	0.026
	hsize				0.404	0.182
	pl030					0.172
Proposed method	db040	0.262	0.019	0.217	0.139	0.160
	age		0.118	0.546	0.716	0.153
	rb090			0.082	0.386	0.026
	hsize				0.405	0.179
	pl030					0.171
Relative differences (in %)	db040	0.283	−2.032	0.000	0.142	−0.073
	age		0.021	0.030	−1.098	−21.220
	rb090			0.339	0.129	−0.580
	hsize				0.108	−1.267
	pl030					−0.445
Method of Münnich et al	db040	0.262	0.019	0.217	0.138	0.160
	age		0.118	0.546	0.715	0.151
	rb090			0.081	0.386	0.026
	hsize				0.366	0.045
	pl030					0.172
Relative differences (in %)	db040	0.190	−1.139	0.110	−0.536	−0.172
	age		0.195	0.037	−1.175	−22.164
	rb090			−0.288	0.233	2.091
	hsize				−9.423	−74.970
	pl030					−0.055

computation time considerably, therefore a reasonable tradeoff has been used. All in all, the correlation structure of the simulated populations is very close to the expected one. For the method of Münnich et al (2003) and Münnich and Schürle (2003), the contingency coefficient of age and citizenship also suffers from using age category as conditioning variable. Moreover, the relationships between household size (`hsize`) and the variables economic status (`pl030`) and citizenship (`pb220a`) are not well reflected. This is because these authors suggest to use only stratum, age category and gender as conditioning variables for the simulation of additional categorical variables. Including household size as conditioning variable in the estimation of the conditional multivariate distributions leads to an improvement of the contingency coefficients (results not shown), but causes another problem. Since the size of the sample is rather small, only 3432 of the possible 51030 combinations of region, age category, gender, household size, economic status and citizenship exist in the sample. Hence the resulting populations cannot contain any other combinations either. Even though many of the combinations that do not occur are structural zeros, such low variation in the population is simply not realistic as

Table 4 Evaluation of personal net income based on the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation. Weighted values from the sample data are compared to average results from 100 simulated populations.

		%Zeros	**5%**	**Median**	**Mean**	**95%**	**SD**
Sample		11.39	2800.00	15428.26	17084.37	36000.00	11589.52
Averages of	MP	11.41	2728.46	15703.79	17135.24	35682.04	11386.47
simulated	TR	11.41	3643.36	14858.96	16980.55	37310.49	11257.74
populations	TN	11.41	4946.36	14949.58	17118.29	36566.82	10296.96
Relative	MP	0.14	−2.55	1.79	0.30	−0.88	−1.75
differences	TR	0.19	30.12	−3.69	−0.61	3.64	−2.86
(in %)	TN	0.19	76.66	−3.10	0.20	1.57	−11.15

it is very likely that there is a significant number of random zeros resulting from the small sample size (see Section 3.3). In short, the proposed method has the advantage that the information from the household size can be included in the simulation of additional variables since the multinomial models allow to simulate combinations that do not exist in the sample.

In Table 4, simulated personal net income is evaluated based on various quantities of interest: the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation. Weighted values from the sample data are compared to the average results from the simulated populations for each of the three investigated methods. The relative differences are again used for evaluation. Clearly, the MP approach performs best with an excellent overall fit. For the two-step linear regression procedures, there is considerable deviation in the 5% quantile (cf. Figure 2, left). Due to the better fit and the more accurate standard deviation, the TR approach may be favorable over the TN approach.

3.3 Influence of sample size and sampling design

In this section, the synthetic population data from Section 3.1 are used to evaluate the effect of different sample sizes and sampling designs on the proposed framework in a simulation study. It may not be optimal to use population data that have been generated with the same methodology for such an analysis, but since real population data are not available, this is the only possible way to investigate these issues.

Concerning the sample size, two different scenarios are considered: (*i*) 6 000 households, which is roughly the real sample size, and (*ii*) 1% samples, which corresponds to about 35 000 households. In addition, the following two sampling designs are investigated, both of which are frequently used for EU-SILC in practice:

1. Stratified simple random sampling of households by region.

2. Stratified simple random sampling of individuals by region. Then all individuals belonging to the same household as any of the sampled individuals are collected and added to the sample.

The sample sizes were in both cases chosen proportional to the strata sizes. This leads to approximately equal weights for the first design, and weights approximately inverse proportional to the corresponding household sizes for the second design. For each combination of sample size and sampling design, 25 samples are drawn from the initial population. Then 10 populations are simulated for each sample, resulting in a total of 250 synthetic populations. Furthermore, calibration using different choices of variables did not have a strong impact on the characteristics of the resulting variables (results not shown). Since households are sampled, however, the resulting sample weights in general do not sum up to the number of individuals in the population. Therefore, calibration on the marginal totals of the regions is performed.

Since the proposed framework allows to simulate combinations of categorical variables that do not occur in the underlying sample, empty cells in the contingency tables are analyzed. Table 5 lists the number of empty cells for the initial population ($\#Initial$), the average percentage of these cells that are no longer empty for the simulated populations (false nonempties, $\%FN$), the average number of of additional random empty cells for the samples introduced by the sampling process ($\#Random$), and the average proportion of these cells that are still empty for the simulated populations (false empties, $\%FE$).

For all scenarios, only a very low percentage of combinations that do not exist in the initial population are introduced in the simulated populations. Note that not all empty cells in the contingency table of the initial population are structural zeros. Just because a certain combination does not occur in a specific population does not mean that it is impossible to occur. Thus new combinations introduced in the simulated populations may very well be realistic. In any case, the probability for generating a combination that is in fact a structural zero is very low due to the low percentage of false nonempties.

On the other hand, the large majority of combinations that randomly do not exist in the corresponding sample due to the sampling process are generated in the synthetic populations. Nevertheless, in particular for the small real

Table 5 Analysis of empty cells in the contingency tables of the categorical variables. 250 simulated populations are evaluated using the number of empty cells for the initial population ($\#Initial$) and the respective average percentage of false nonempties ($\%FN$), as well as the average number of of additional random empty cells for the samples ($\#Random$) and the respective average proportion of false empties ($\%FE$).

Size	Design	#Initial	%FN	#Random	%FE
Real	1	37730	0.61	10006.48	33.80
Real	2	37730	0.63	9540.84	29.59
1%	1	37730	0.99	6782.76	9.29
1%	2	37730	0.99	6327.52	9.74

sample size, there is a considerable amount of such combinations that still do not occur in the simulated populations. The main reason for this is that large households do not occur very frequently in the initial population, hence there is only a low number of such households in the samples, which in turn makes it difficult to reproduce the full variation of possible combinations. This also explains why the first scenario with the real sample size and simple random sampling of households leads to the largest proportion of false empties, as it results in the lowest expected absolute frequencies of large households. To summarize, considering the small sample size for the first two scenarios and the resulting large number of random empty cells in the contingency tables for the samples, the proposed procedure performs quite well.

In Table 6, the contingency coefficients between the categorical variables from the initial population are compared to the average results from the simulated populations for each of the four sampling scenarios. For the real sample size, there are considerable differences specifically in the contingency coefficients between the variables region (`db040`), age and gender (`rb090`). This is because the household structure is simulated by resampling households from the sample, which due to the small size does not account for all the variation in the initial population. However, since the dependencies within a household are highly complex, the results with the simple resampling approach can still be considered very reasonable. In addition, most of the other relationships are very well reflected. The 1% samples are of course much less affected by the effect of resampling households, and all in all the results are excellent.

Table 7 contains an evaluation of the simulated personal net income based on the following quantities of interest: the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation. It should be noted that the reference values for the initial population are computed from the income generated by the MP approach, since this this gave the best fit compared to the original sample data (see Section 3.1). The results do not suggest a very strong influence of the sample size or the sampling design and are similar to those from the comparison to the original sample data in Section 3.2. For the real sample size, only a small effect of the sampling design on the percentage of zeros is visible in all methods. Furthermore, the sampling design appears to have a slight impact on the two-step linear regression methods in general, most notably on the 5% and 95% quantiles and the standard deviation. In any case, the MP approach clearly gives excellent results and performs best, while the TR method is favorable over the TN method for the two-step approach.

4 Conclusions

This paper introduced a flexible framework for simulating population data for household surveys based on available sample data, which is implemented along with diagnostic plots in the R package `simPopulation`. No auxiliary information is used in the procedure, and stratification allows to account for heterogeneities such as regional differences.

Table 6 Pairwise contingency coefficients of the categorical variables for the initial population, as well as average results from 250 simulated populations for each of the four sampling scenarios.

		age	rb090	hsize	pl030	pb220a
Population	db040	0.261	0.020	0.217	0.138	0.160
	age		0.118	0.546	0.716	0.153
	rb090			0.082	0.386	0.026
	hsize				0.405	0.179
	pl030					0.172
Real size, Design 1	db040	0.337	0.025	0.256	0.153	0.162
	age		0.141	0.565	0.717	0.162
	rb090			0.086	0.387	0.029
	hsize				0.408	0.186
	pl030					0.174
Relative differences (in %)	db040	29.055	28.335	18.069	10.265	0.740
	age		19.209	3.512	0.209	6.091
	rb090			5.692	0.355	8.711
	hsize				0.835	3.952
	pl030					1.316
Real size Design 2	db040	0.347	0.028	0.239	0.157	0.165
	age		0.142	0.560	0.716	0.162
	rb090			0.080	0.388	0.027
	hsize				0.404	0.188
	pl030					0.176
Relative differences (in %)	db040	32.810	41.592	10.129	13.667	3.091
	age		20.100	2.631	0.121	6.142
	rb090			−1.475	0.592	2.197
	hsize				−0.097	4.590
	pl030					2.664
1% sample, Design 1	db040	0.278	0.020	0.223	0.141	0.161
	age		0.123	0.549	0.716	0.153
	rb090			0.082	0.385	0.027
	hsize				0.406	0.182
	pl030					0.171
Relative differences (in %)	db040	6.352	1.480	2.951	1.640	0.577
	age		4.048	0.611	0.029	0.524
	rb090			0.532	−0.182	1.185
	hsize				0.308	1.212
	pl030					−0.062
1% sample, Design 2	db040	0.277	0.021	0.221	0.140	0.161
	age		0.122	0.549	0.716	0.154
	rb090			0.082	0.386	0.026
	hsize				0.406	0.179
	pl030					0.171
Relative differences (in %)	db040	6.024	4.893	1.775	1.067	0.207
	age		3.720	0.593	0.020	0.755
	rb090			0.101	0.138	0.559
	hsize				0.196	−0.065
	pl030					−0.525

Table 7 Evaluation of personal net income based on the percentage of zeros, 5% quantile, median, mean, 95% quantile and standard deviation. Values from the initial population are compared to average results from 250 simulated populations for each of the four sampling scenarios.

		%Zeros	**5%**	**Median**	**Mean**	**95%**	**SD**
Population		11.40	2719.73	15700.65	17130.72	35677.48	11390.32
Real size,	MP	11.26	2669.72	15667.54	17064.57	35601.94	11232.09
Design 1	TR	11.28	3514.98	14770.39	16900.82	37305.08	11209.41
	TN	11.28	4755.06	14990.06	17377.93	38028.43	10935.30
Relative	MP	−1.15	−1.84	−0.21	−0.39	−0.21	−1.39
differences	TR	−1.00	29.24	−5.92	−1.34	4.56	−1.59
(in %)	TN	−1.00	74.84	−4.53	1.44	6.59	−3.99
Real size,	MP	11.44	2708.81	15665.23	17136.36	35826.75	11385.18
Design 2	TR	11.45	3397.59	14879.76	17097.54	38101.73	11540.92
	TN	11.45	4743.06	15095.90	17610.83	38929.09	11278.78
Relative	MP	0.38	−0.40	−0.23	0.03	0.42	−0.05
differences	TR	0.45	24.92	−5.23	−0.19	6.79	1.32
(in %)	TN	0.45	74.39	−3.85	2.80	9.11	−0.98
1% sample,	MP	11.37	2720.79	15695.71	17113.44	35643.94	11279.51
Design 1	TR	11.37	3529.14	14834.87	16948.50	37324.55	11196.36
	TN	11.37	4838.43	15049.16	17434.75	38057.47	10907.39
Relative	MP	−0.22	0.04	−0.03	−0.10	−0.09	−0.97
differences	TR	−0.19	29.76	−5.51	−1.06	4.62	−1.70
(in %)	TN	−0.19	77.90	−4.15	1.77	6.67	−4.24
1% sample,	MP	11.36	2723.38	15699.97	17134.05	35661.53	11340.96
Design 2	TR	11.37	3406.91	14913.92	17085.76	37937.72	11455.55
	TN	11.37	4810.99	15134.27	17628.18	38840.56	11211.61
Relative	MP	−0.31	0.13	−0.00	0.02	−0.04	−0.43
differences	TR	−0.25	25.27	−5.01	−0.26	6.34	0.57
(in %)	TN	−0.25	76.89	−3.61	2.90	8.87	−1.57

The proposed framework is applicable to a broad class of surveys and led to excellent results in an application to EU-SILC. For simulation of personal net income, using multinomial models combined with random draws from the resulting categories and generalized Pareto tail modeling performed better than two-step regression, but is computationally more expensive. The computation time of the multinomial models thereby strongly depends on the number of categories used in the discretization. Concerning the two-step approach, trimming combined with random draws from the residuals appeared to be favorable. Nevertheless, the choice of method also depends on the purpose. For simulation studies in survey statistics, it is important not to favor any of the investigated methods by the underlying data generation procedure in order to avoid biased simulation results.

Acknowledgements The authors are grateful to the referees for helpful comments and suggestions.

References

Alfons A (2010) `simFrame`: Simulation framework. R package version 0.3.7

Alfons A, Kraft S (2010) `simPopulation`: Simulation of synthetic populations for surveys based on sample data. R package version 0.2.1

Alfons A, Templ M, Filzmoser P (2010a) An object-oriented framework for statistical simulation: The R package `simFrame`. Journal of Statistical Software 37(3):1–36

Alfons A, Templ M, Filzmoser P (2010b) Simulation of EU-SILC population data: Using the R package `simPopulation`. Research Report CS-2010-5, Department of Statistics and Probability Theory, Vienna University of Technology

Atkinson T, Cantillon B, Marlier E, Nolan B (2002) Social Indicators: The EU and Social Inclusion. Oxford University Press, New York, ISBN 0-19-925349-8

Clarke G (1996) Microsimulation: an introduction. In: Clarke G (ed) Microsimulation for Urban and Regional Policy Analysis, Pion, London

Drechsler J, Bender S, Rässler S (2008) Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. Trans Data Priv 1(3):105–130

Embrechts P, Klüppelberg G, Mikosch T (1997) Modelling Extremal Events for Insurance and Finance. Springer, New York, ISBN 3-540-60931-8

Eurostat (2004) Description of target variables: Cross-sectional and longitudinal. EU-SILC 065/04, Eurostat, Luxembourg

Horvitz D, Thompson D (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47(260):663–685

Kendall M, Stuart A (1967) The Advanced Theory of Statistics, vol 2, 2nd edn. Charles Griffin & Co. Ltd., London

Kleiber C, Kotz S (2003) Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, Hoboken, ISBN 0-471-15064-9

Kraft S (2009) Simulation of a Population for the European Living and Income Conditions Survey. Master's thesis, Vienna University of Technology

Meyer D, Zeileis A, Hornik K (2006) The `strucplot` framework: Visualizing multi-way contingency tables with `vcd`. J Stat Softw 17(3):1–48

Meyer D, Zeileis A, Hornik K (2010) `vcd`: Visualizing Categorical Data. R package version 1.2-9

Münnich R, Schürle J (2003) On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen

Münnich R, Schürle J, Bihler W, Boonstra HJ, Knotterus P, Nieuwenbroek N, Haslinger A, Laaksonen S, Eckmair D, Quatember A, Wagner H, Renfer JP, Oetliker U, Wiegert R (2003) Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen

Raghunathan T, Reiter J, Rubin D (2003) Multiple imputation for statistical disclosure limitation. J Off Stat 19(1):1–16

R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0

Reiter J (2009) Using multiple imputation to integrate and disseminate confidential microdata. Int Stat Rev 77(2):179–195

Rubin D (1993) Discussion: Statistical disclosure limitation. J Off Stat 9(2):461–468

Sarkar D (2008) Lattice: Multivariate Data Visualization with R. Springer, New York, ISBN 978-0-387-75968-5

Sarkar D (2011) `lattice`: Lattice Graphics. R package version 0.19-17

Simonoff J (2003) Analyzing Categorical Data. Springer, New York, ISBN 0-387-00749-0

Templ M, Alfons A (2010) Disclosure risk of synthetic population data with application in the case of EU-SILC. In: Domingo-Ferrer J, Magkos E (eds) Privacy in Statistical

Databases, Lecture Notes in Computer Science, vol 6344, Springer, Heidelberg, pp 174–186

Walker A (1977) An efficient method for generating discrete random variables with general distributions. ACM Trans Math Softw 3(3):253–256

Weisberg S (2005) Applied Linear Regression, 3rd edn. Wiley, Hoboken, ISBN 0-471-66379-4