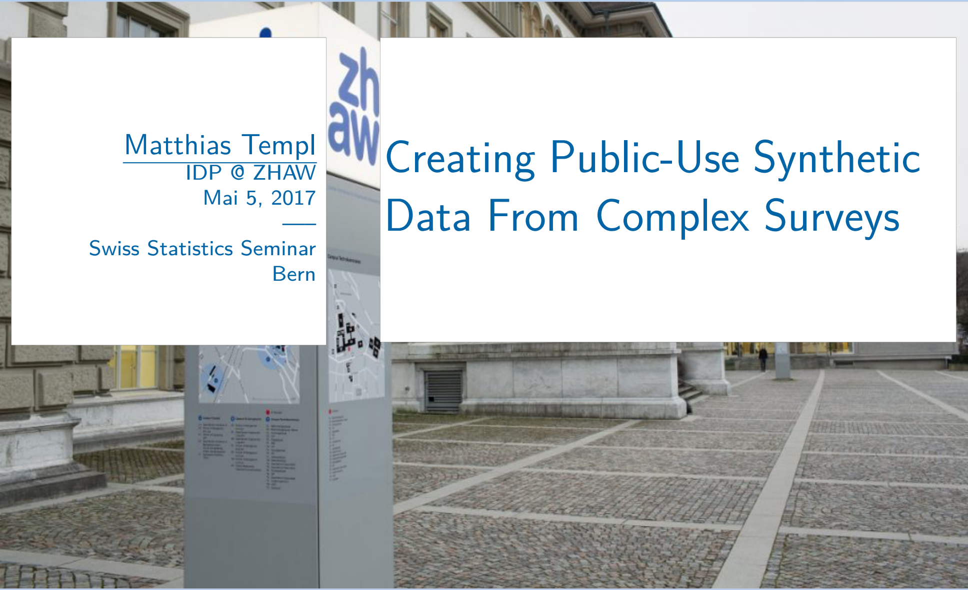
Creating Public-Use Synthetic Data From Complex Surveys

Matthias Templ
IDP @ ZHAW

Mai 5, 2017

Swiss Statistics Seminar
Bern



Overview

Motivation

What are close-to-reality data?

Model-based simulation methods

Package simPop

Example: Simulating the Austrian Population from the European Statistics on Income and Living Conditions Survey (simplified version)

Preconception

Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering

Unluckily, often some preconceptions are present related to synthetic data, some of them:

- We have a lot of data and do not need synthetic data/populations
- Others don't work with synthetic data
- Synthetic data are not real/true data
- Synthetic data $\rightarrow$ credibility loss
- We have more important issues to do
- Just a hobby from science in a dreaming spire

"New opinions are always suspected, and usually opposed, without any other reason but because they are not already common".
(John Locke, 1689)

Preconception

Unluckily, often some preconceptions are present related to synthetic data, some of them:

- We have a lot of data and do not need synthetic data/populations
- Others don't work with synthetic data
- Synthetic data are not real/true data
- Synthetic data $\rightarrow$ credibility loss
- We have more important issues to do
- Just a hobby from science in a dreaming spire

"New opinions are always suspected, and usually opposed, without any other reason but because they are not already common".
(John Locke, 1689)

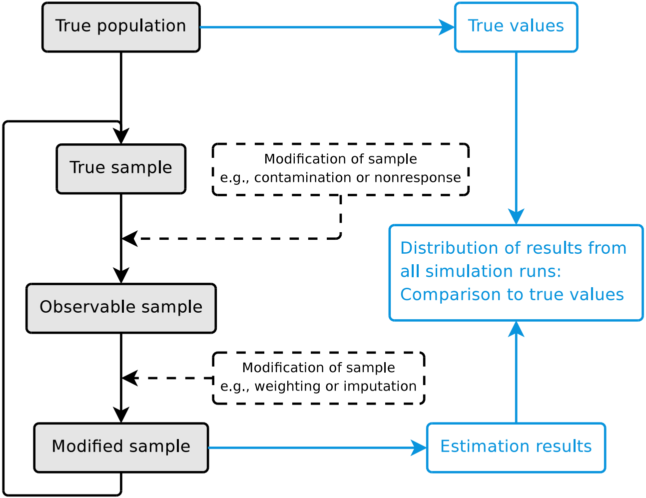
Why synthetic populations?

1) **Augmentation** of data/populations with interesting variables from different sources
2) **Simulation studies** for the evaluation and development of methods
    - complex (design-based) simulation studies in survey methodology
    - influence of sampling designs on methods and results

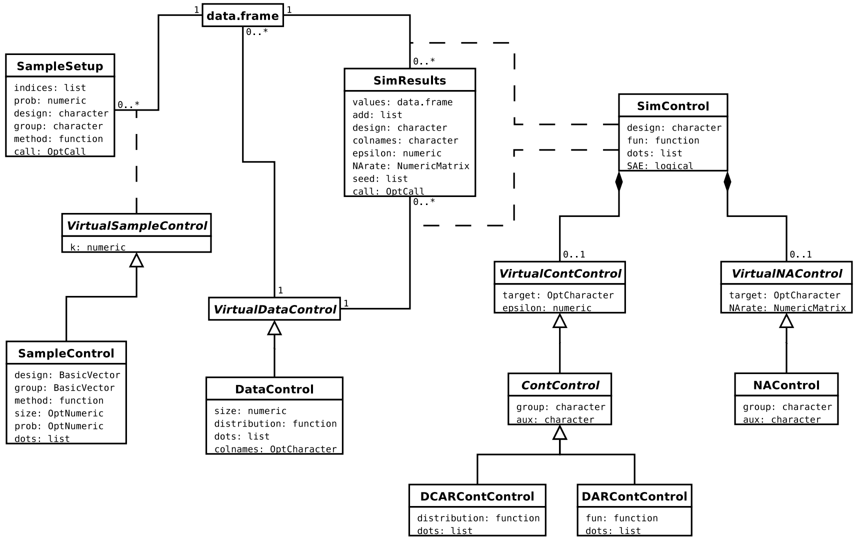
Why synthetic populations?

1) **Augmentation** of data/populations with interesting variables from different sources
2) **Simulation studies** for the evaluation and development of methods
   - complex (design-based) simulation studies in survey methodology
   - influence of sampling designs on methods and results

ad 2) design-based **simulation studies**



ad 2) ... in software (simFrame)



Why synthetic populations?

3) for agent-based- and/or **micro-simulation**
   - e.g. health planning, spread of diseases, climate change forecasting, forecasting demographic and economic changes – all on individual (micro-level) basis
   - Starting point is a population of all individuals at time T_0
   - Hot topic in research. "Loved" by manages and ecometricians.

4) **Public-use data** for research and education

5) because the **disclosure risk** $\rightarrow 0$ (confidentiality issues ✓) (Templ and Alfons 2010)

Remark: one can always draw a sample from a population.

Motivation

What are close-to-reality data?

Model-based simulation methods

Package simPop

Example: Simulating the Austrian Population from the European Statistics on Income and Living Conditions Survey (simplified version)

Close-to-reality data

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- actual sizes of regions and strata need to be reflected
- marginal distributions and interactions between variables should be represented correctly
- hierarchical and cluster structures have to be preserved
- sometimes some marginal distributions must exactly match known values
- data confidentiality must be ensured
  - no replication of units
  - avoid to use commonly used (model-based) imputation methods

First applications (historically)

- Clarke, Clarke, Birkin, Rees und Wilson (1984) simulated a population from aggregated data for the British (**health**) care organisation.
- Estimation of the **demand** of water (Clarke and Holm, 1987; Williamson, Birkin and Rees, 1998)
- From 1998 onwards a lot of applications such as
  - **health planning** (Brown and Harding, 2002; Tomintz, Clarke, and Rigby, 2008), (Smith, Pearce, and Harland, 2011),
  - **transportation** (Beckman, Baggerly, and McKay, 1996; Barthelemy and Toint, 2013)
  - **environmental planning** (Williamson, 2002).
- Evaluation and comparison of estimators and methods in DACSEIS, EUREDIT, AMELI, ..., research projects European level

Motivation

What are close-to-reality data?

Model-based simulation methods

Package simPop

Example: Simulating the Austrian Population from the European Statistics on Income and Living Conditions Survey (simplified version)

Too simplistic:

-
```
mvtnorm::rmvnorm(n = 500,
                 mean = c(1,2),
                 sigma = matrix(c(4,2,2,3), ncol=2))
```
-
```
X <- T %*% t(B) + E # component model
```
- simple model-based
- too simulate data with the help of model-based imputation methods suggested by Rubin (1993) and many other authors
- use of simple sampling methods
- using Copulas to simulate multivariate data
- machine and deep learning methods

For complex data these methods are too simplistic

Model-based methods

... for the simulation of close-to-reality populations

- **multi-phase** process and sequential process
- special use of **regression** methods

Additionally needed

- **Calibration methods** to calibrate on known population characteristics
- **Combinatorial optimization methods** for the calibration of populations
- Tools to deal with special data problems, such as
  - **heaping** (e.g. age heaping)
  - **imputation methods** for missing values

Model-based methods

... for the simulation of close-to-reality populations

- **multi-phase** process and sequential process
- special use of **regression** methods

Additionally needed

- **Calibration methods** to calibrate on known population characteristics
- **Combinatorial optimization methods** for the calibration of populations
- Tools to deal with special data problems, such as
  - **heaping** (e.g. age heaping)
  - **imputation methods** for missing values

simPop history

- Theory: EU-FP7 project *AMELI* (Templ et.al, 2011)
- Software **simPopulation** (depricated)
- Software **simPop** (Templ, Kowarik, and Meindl 2016a):
  - *Methods and tools for the generation of synthetic populations* (World Bank Project-No. 1129231)
  - *Synthetic populations and microsimulation* (World Bank Project-No 7177468)

Motivation

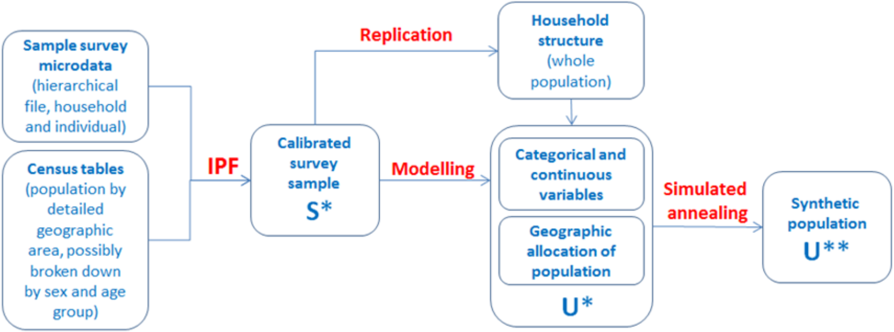
What are close-to-reality data?

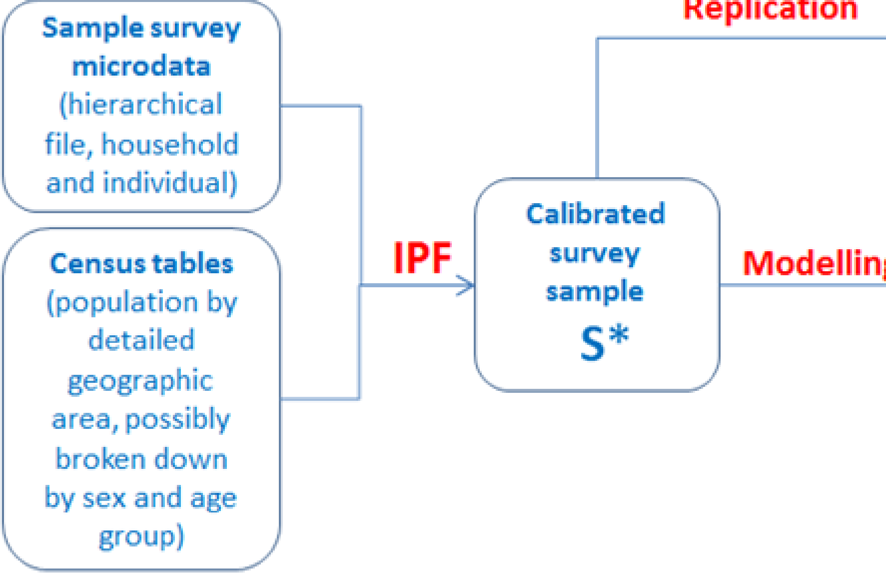
Model-based simulation methods

Package simPop

Example: Simulating the Austrian Population from the European Statistics on Income and Living Conditions Survey (simplified version)

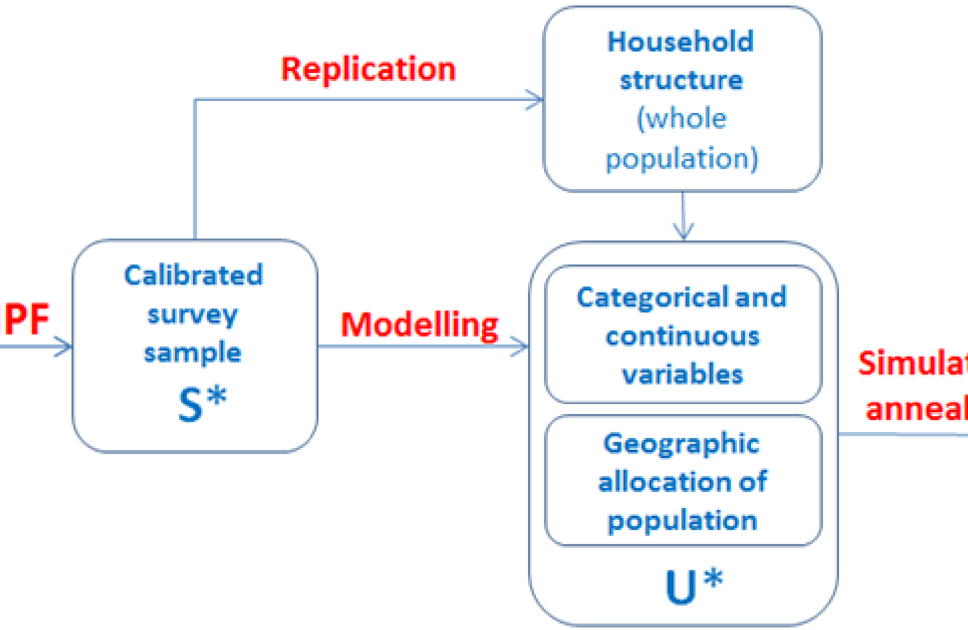
Model-based approach: simplified workflow




Sample survey microdata (hierarchical file, household and individual)
Census tables (population by detailed geographic area, possibly broken down by sex and age group)
IPF
Calibrated survey sample S*
Replication
Modellin

Calibration of a survey sample

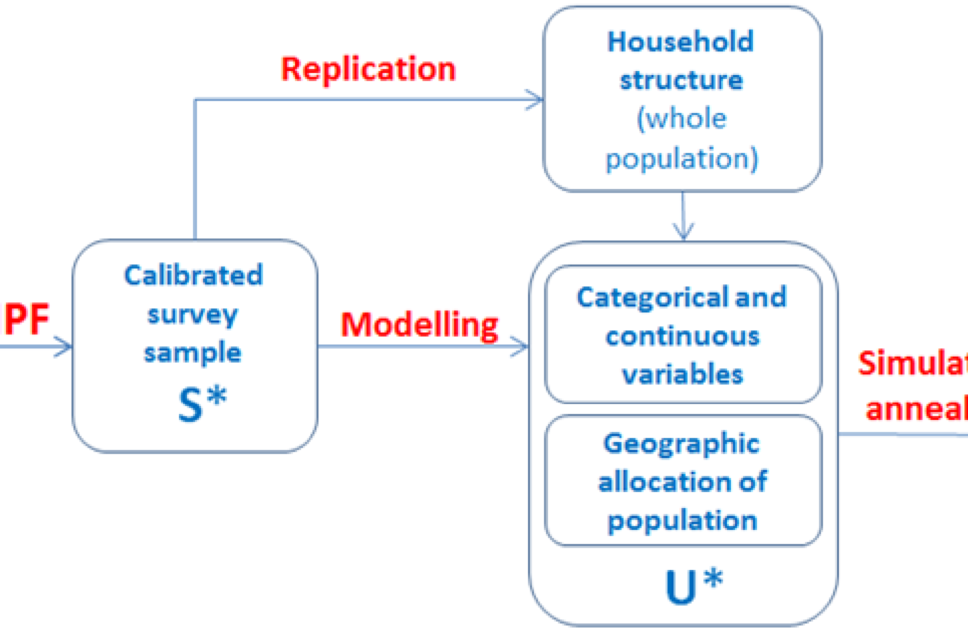
- $S_i = 1$ if i in sample, otherwise 0
- $Y = \sum_{i=1}^{N} y_i$ **unknown** population total with N observations. From sample, estimation with Horwitz-Thompson estimator $\hat{Y}_d = \sum\limits_{i:S_i=1} d_i y_i$, with design-weights $d_i = 1/\pi_i$
- Auxiliary variable $\boldsymbol{x}$ with **known** total $X = \sum_{i=1}^{N} x_i$, and $\sum_{i:S_i=1} d_i x_i \neq X$. Find new weights w_i with $\hat{Y}_w = \sum_{i:S_i=1} w_i y_i$ for that $\sum_{i:S_i=1} w_i x_i = X$ holds, and $\sum_{i:S_i=1} w_i = N$.
- More than one auxiliary variable possible
- Solution: iterative proportional fitting methods (**raking**)
- Cluster-Structures (e.g. persons in households), **iterative proportional updating**


PF
Replication
Household structure (whole population)
Calibrated survey sample
S*
Modelling
Categorical and continuous variables
Geographic allocation of population
U*
Simula
anneal

Replication - model-based: Structure

Example: persons in households data

- **household structure** (core-variables): independently for every combination of household size and strata
- number of households: Horwitz-Thompson estimation
- for confidentiality issues, use only few variables for the structure
- e.g. age $\times$ region $\times$ gender ($\forall$ strata & households)


Replication
Household structure (whole population)
PF
Calibrated survey sample
S*
Modelling
Categorical and continuous variables
Geographic allocation of population
U*
Simula
anneal

Model-based approach: workflow

After setting up the household structure, additional variables are simulated using regression models:

1. simulation of categorical variables
2. simulation of (semi-) continuous variables
3. (simulation of compositions, e.g. income variables)

- stratification to reflect heterogeneity
- account for sampling weights
- account for missing values

Estimation on the sample (sketch)

$$
\textit{sample} \quad \boldsymbol{S} = \begin{pmatrix} \overbrace{x_{1,1} & x_{1,2} & \cdots & x_{1,j}}^{\text{“predictors”}} & \overbrace{x_{1,j+1}}^{\text{response}} & \overbrace{x_{1,j+2} & \cdots}^{\text{rest}} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & x_{2,j+1} & x_{2,j+2} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & x_{n,j+1} & x_{n,j+2} & \cdots \end{pmatrix}
$$

- design matrix to model $\boldsymbol{x}_{j+1}$
- Models of any complexity can be specified for each variable.
- estimation of the parameters, the $\boldsymbol{\beta}$'s, using multinomial regression, naive bayes, 2-step-approaches, regression trees, ctrees, ...

Prediction of the population (sketch)

$$population \quad \boldsymbol{U} = \begin{pmatrix} \overbrace{\begin{matrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \cdots & \hat{x}_{1,j} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \cdots & \hat{x}_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{n,2} & \cdots & \hat{x}_{N,j} \end{matrix}}^{\hat{\boldsymbol{\beta}} \times \text{“pred.”} \approx} & \overbrace{\begin{matrix} \hat{x}_{1,j+1} \\ \hat{x}_{1,j+1} \\ \vdots \\ \vdots \\ \hat{x}_{1,j+1} \end{matrix}}^{\hat{\mathbf{x}}_{j+1}} \end{pmatrix}$$

- We don't take the expected values, but draw from predictive distributions.

Model-based simulation of new variables

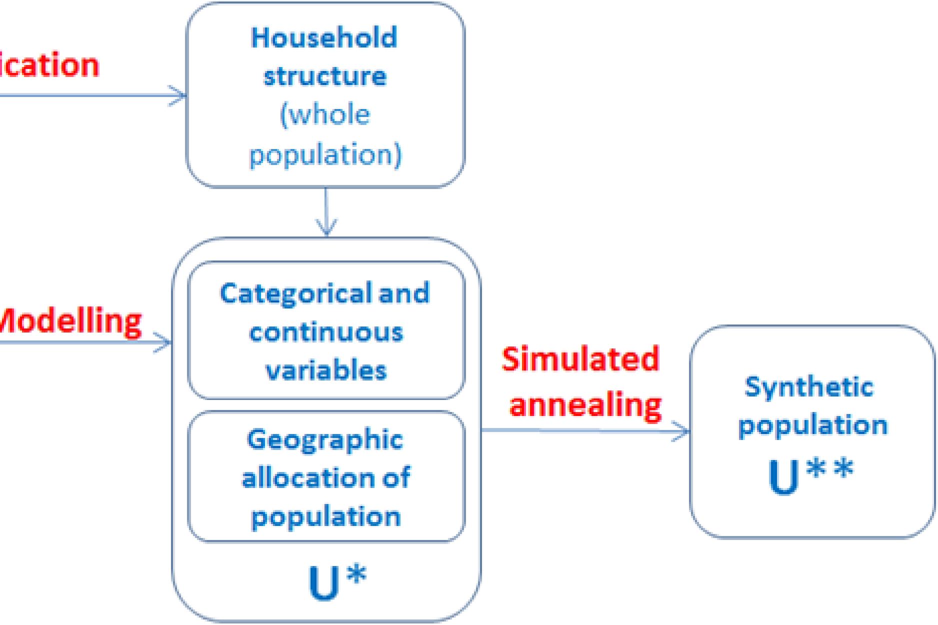
Categorical variables:

- methods: multinomial regression, naive bayes, 2-step-approaches, regression trees, ctrees, ...

Continuous or semi-continuous variables

- multinomial model and draw from categories
- robust or ordinary least squares methods, glm's
- two-step approach for semi-continuous variables

Random noise is added by draws from the residuals or from the (normal) distribution of the residuals


ication
Household structure (whole population)
Modelling
Categorical and continuous variables
Geographic allocation of population
U*
Simulated annealing
Synthetic population U**

Combinatorial optimization

These techniques can be used to

- calibrate synthetic populations to receive consistent estimates for known marginal distributions (*swapping*, *target swapping*)
- *add finer geographical levels*
- Methods: **simulated annealing**, genetic algorithms, ...

Untold stories

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- adding finer geographical information
- age heaping
- imputation of item non-responses within the procedures
- evaluation of the disclosure risk and quality/utility of the synthetic population

Motivation

What are close-to-reality data?

Model-based simulation methods

Package simPop

Example: Simulating the Austrian Population from the European Statistics on Income and Living Conditions Survey (simplified version)

R package simPop

- all mentioned methods & (much) more
- strictly object-Oriented (S4 class implementation)
- efficiently programmed, can be used for huge data sets
- parallel computing is applied automatically
- "documentation" accepted on December 2015 in the Journal of Statistical Software
- last developments were supported by funds from the World bank

EU-SILC

Using real-world data and simulating about 500 variables for several countries Templ, Spiess, Bergeat, and Meindl (2016b), Bergeat, Templ, and Spiess (2016)

```
library(simPop) # call simPop in R
# number of persons
nrow(origData)

## [1] 11725

# number of households (household ID: db030):
uniqueN(origData$db030)

## [1] 4641
```

specifyInput()

```
inp <- specifyInput(origData,          # sample survey
                    hhid = "db030",
                    hhsize = "hsize",
                    strata = "db040", # by region
                    weight = "rb050"); inp
```

```
##
##  --------------
## survey sample of size 11725 x 19
##
##  Selected important variables:
##
##  household ID: db030
##  personal ID: pid
##  variable household size: hsize
##  sampling weight: rb050
##  strata: db040
```

Calibration of the survey sample

```r
data("totalsRG"); data("totalsRGtab")
totalsRGtab
```

```
##          db040
## rb090    Burgenland Carinthia Lower Austria Salzburg Styria  Tyrol
##   female     146980    285797        828087   722883 274675 619404
##   male       140436    270084        797398   702539 259595 595842
##          db040
## rb090     Upper Austria Vienna Vorarlberg
##   female         368128 916150     190343
##   male           353910 850596     184939
```

Calibration:

```r
addWeights(inp) <- calibSample(inp, totalsRG)
```

Simulating the structure

```
synthP <- simStructure(inp,
                       method = "direct",
                       basicHHvars = c("age", "rb090", "db040"))
synthP
```

```
##
## --------------
## synthetic population  of size
##  8504755 x 7
##
## build from a sample of size
## 11725 x 19
## --------------
##
## variables in the population:
## db030,hsize,age,rb090,db040,pid,weight
```

Categorical variables

```
synthP <- simCategorical(synthP,
             regModel = "available", # or formulas
             additional = c("pl030", "pb220a"),
             method = "multinom") # ctree, randomForest, ...
synthP
```

```
##
## --------------
## synthetic population  of size
##  8504755 x 9
##
## build from a sample of size
## 11725 x 19
## --------------
##
## variables in the population:
## db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a
```

Continuous variables

```
synthP <- simContinuous(synthP,
              additional = "netIncome",
              regModel = ~ rb090 + hsize + pl030 + pb220a)
synthP
```

```
##
## --------------
## synthetic population  of size
##  8504755 x 11
##
## build from a sample of size
## 11725 x 19
## --------------
##
## variables in the population:
## db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a,netIncomeCat,ne
```

Calibrate population

- Again: census information to calibrate. External information (n-dimensional table) is available, e.g marginals on region gender economic status.
- We add these marginals to the object and calibrate afterwards

```
synthP <- addKnownMargins(synthP, totalsRG)
```

```
synthP <- calibPop(synthP)
```

```
  as also true for other functions, many parameters
available, here optional:  split="db040", temp=1, eps.facto
maxiter=200, temp.cooldown=0.975, factor.cooldown=0.85,
min.temp=0.001, verbose=FALSE
```

Quality and disclosure risk of the population

Many utility measures possible, from **simple indicators**, to **visual comparisons**, to compare **point** and **variance estimates** for indicators, compare results from **models**.

The aim is always to **compare the sample information** or the information on known characteristics with results from the **synthetic population**.

- disclosure risk, see Templ and Alfons (2010)
- utility:
  - quality indicators: Templ (2017, 2015)
  - population based on EU-SILC: Alfons, Kraft, Templ, and Filzmoser (2011b), Bergeat et al. (2016), for employer-employee data: Templ and Filzmoser (2014)

We show two visual comparisons (can be done on finer detail)

Quality/utility of the population

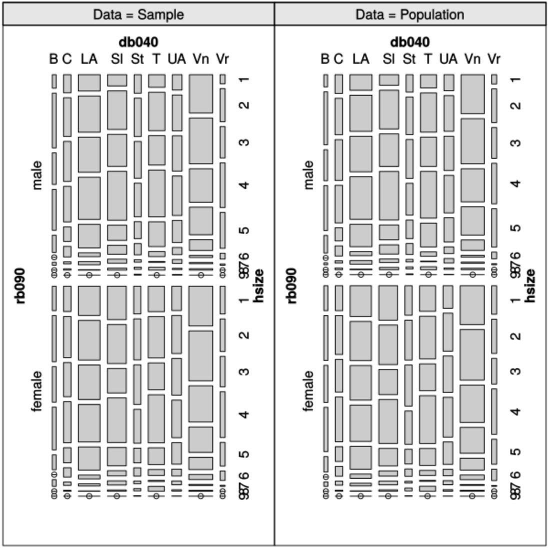
Tables (HT-(weighted) estimation):

```
tab <- spTable(synthP,
        select=c("rb090", "db040", "hsize"))
```

Show frequencies visually:

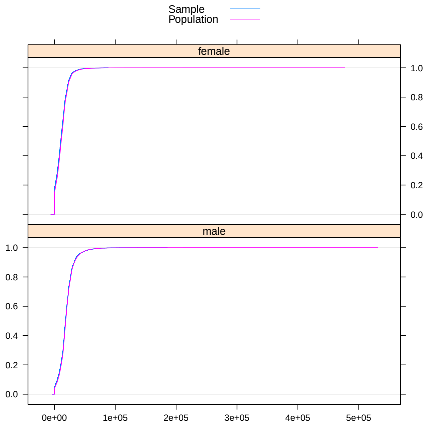
```
spMosaic(tab, # method = "color",
  labeling = labeling_border(
             abbreviate = c(db040 = TRUE)))
```

Quality/utility of the population



Quality/utility of the population

```
spCdfplot(synthP,
          x = "netIncome", cond="rb090", layout=c(1,2))
```



Conclusions

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be very precisely estimated
- the synthetic populations are confidential
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest) can be used
- can also work as input for microsimulation or design-based simulation studies
- open-access, public-use data

References

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to eu-silc. Statistical Methods & Applications, 20(3):383–407, 2011a.

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods & Applications, 20(3):383–407, 2011b.

M. Bergeat, M. Templ, and L. Spiess. Public use files for EU-SILC – utility analysis. SGA PUF Deliverable D3.2, Statistics Austria, 2016.

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

M. Templ. Quality indicators for statistical disclosure methods: A case study on the structure of earnings survey. Journal of Official Statistics, 31(4):737–761, 2015.

M. Templ. Statistical Disclosure Control for Microdata. Methods and Applications in R. Springer, New York, 2017.

M. Templ and A. Alfons. Disclosure risk of synthetic population data with application in the case of EU-SILC. In J. Domingo-Ferrer and E. Magkos, editors, Privacy in Statistical Databases, volume 6344 of Lecture Notes in Computer Science, pages 174–186. Springer, Heidelberg, 2010.

M. Templ and P. Filzmoser. Simulation and quality of a synthetic close-to-reality employer–employee population. Journal of Applied Statistics, 41(5):1053–1072, 2014.

M. Templ, A. Kowarik, and B. Meindl. Simulation of synthetic complex data: The R-package simPop. Journal of Statistical Software, pages 1–39, 2016a. accepted for publication.

M. Templ, L. Spiess, M. Bergeat, and B. Meindl. Public use files for EU-SILC. SGA PUF Deliverable D3.1, Statistics Austria, 2016b.