Simulation of public-use files from complex survey and population data

Matthias Templ (ZHAW Winterthur), Alexander Kowarik (Statistics Austria)

July 2017

Why synthetic populations?

- **comparison of methods**, e.g. in design-based simulation studies
- **policy modelling** on individual level (e.g health planning, climate change, demographic change, economic change, ...)
- **teaching** (e.g. teaching of survey methods)
- creation of public-/scientific-use files with (very) **low disclosure risk**
- data availability is often a problem (legal issues, costs,...)

Remark: We always can draw samples from a population. To generate a population is a more general approach.

Properties of close-to-reality data

- 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
- data confidentiality must be ensured
- pure replication of units from the underlying sample should be avoided
- sometimes some marginal distributions must exactly match known values
- calibration: certain marginal distributions should be exactly the same as known from other data sources

Available information

- choice of methods depends on available information:
    - census
    - survey samples
    - aggregated information from samples
    - known marginal distributions from population

Model-based approach

- In general, the procedure consists of four steps:
- setup of the household structure (with additional variables)
- simulation of categorical variables
- simulation of continuous variables
- the splitting continuous variables into components
- Stratification: allows to account for heterogenities (e.g. regional differences)

Model-based approach - the basic structure file

- **direct**: estimation of the population totals for each combination of stratum and household size using the Horvitz-Thompson estimator
- **multinom**: estimation of the conditional probabilities within the strata using a multinomial log-linear model and random draws from the resulting distributions
- **distribution**: random draws from the observed conditional distributions within the strata

Example of variables spanning the basic structure: age × region × sex
($\forall$ strata & households)

Model-based approach - fitting

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

$\longrightarrow$ design matrix to model $\boldsymbol{x}_{j+1}$ (account for interactions, etc.).

$\longrightarrow$ estimation of the $\boldsymbol{\beta}$'s

Model-based approach - prediction

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

we don't took expected values but draw from predictive distributions

Model-based approach - categorical variables

Estimation of the β's

- **multinom**: estimation of the conditional probabilities using multinomial log-linear models and random draws from the resulting distributions. Can deal with structural zeros.
- **distribution**: random draws from the observed conditional distributions of their multivariate realizations
- **ctree**: for using classification trees
- **ranger**: for using random forest

simCategorical()

Model-based approach - continuous variables

Similar to the categorical case, but models differ.

- **multinom**: categorize first, then draw from the predictive distributions
- **lm**: for using (two-step) regression models combined with random error terms
- **glm's**, e.g. **poisson** for using Poisson regression for count variables
- robust methods
- **ranger**: for using random forest

simContinuous()

Model-based approach - more methods

Components:

- by resampling fractions from survey data (**simComponents()**)

Relations:

- taking relationships between household members into account (**simRelation()**)

Spatial:

- generation of smaller regions given an existing spatial variable and a table (**simSpatialInit()**)

R package simPop

- Templ, Kowarik, and Meindl (2017), Journal of Statistical Software (accepted)
- latest version on CRAN
- development on github
- parallel computing is applied automatically
- efficient implementation

Define the structure

Create an object of class *dataObj* with function **specifyInput()**.

```
inp <- specifyInput(data=origData,
                    hhid="db030",
                    hhsize="hsize",
                    strata="db040",
                    weight="rb050")
```

```
class(inp)
```

```
## [1] "dataObj"
## attr(,"package")
## [1] "simPop"
```

Simulating the basic structural variables

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

```
## [1] "simPopObj"
## attr(,"package")
## [1] "simPop"
```

- output object (*"synthP"*) is of class *simPopObj*
- various functions can be applied to such objects

Simulation of categorical variables

```
synthP <- simCategorical(synthP, additional=c("pl030", "pb220a"),
  method="multinom")
synthP
```

```
##
## --
## synthetic population  of size
##  8182010 x 9
##
## build from a sample of size
## 11725 x 19
## --
##
## variables in the population:
## db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a
```

almost the same for *simContinuous()*

Census information to calibrate

- We add these marginals to the object and calibrate afterwards

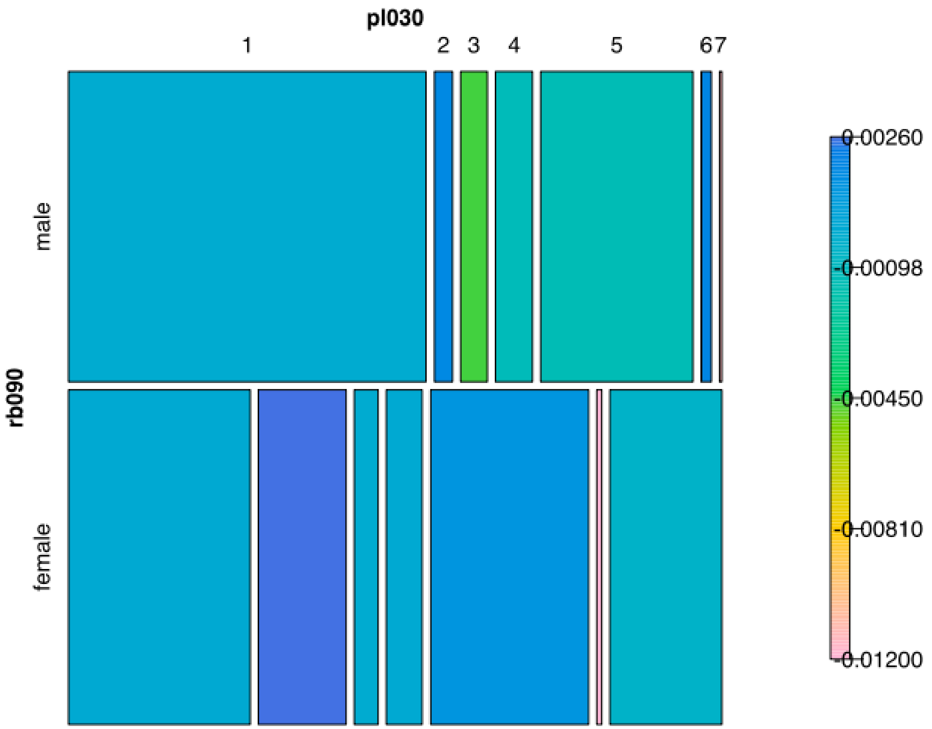
```
synthP <- addKnownMargins(synthP, margins) # add margins
```

```
# calibration using simulated annealing
synthPadj <- calibPop(synthP, split="db040", temp=1,
                      eps.factor=0.00005, maxiter=200,
                      temp.cooldown=0.975,
                      factor.cooldown=0.85,
                      min.temp=0.001, verbose=TRUE)
```

Now: margins of the sample **equals known margins of the population** (not shown here, long computation time.)

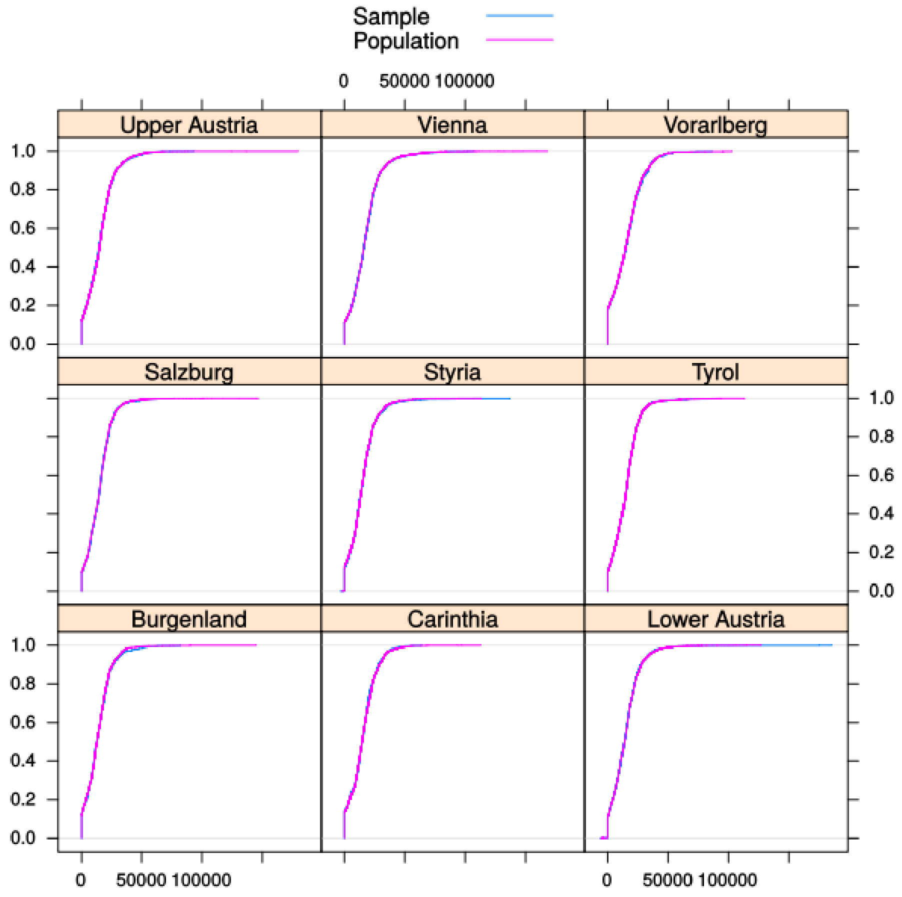
Results

```
tab <- spTable(synthP, select = c("rb090", "pl030"))
spMosaic(tab, method = "color")
```



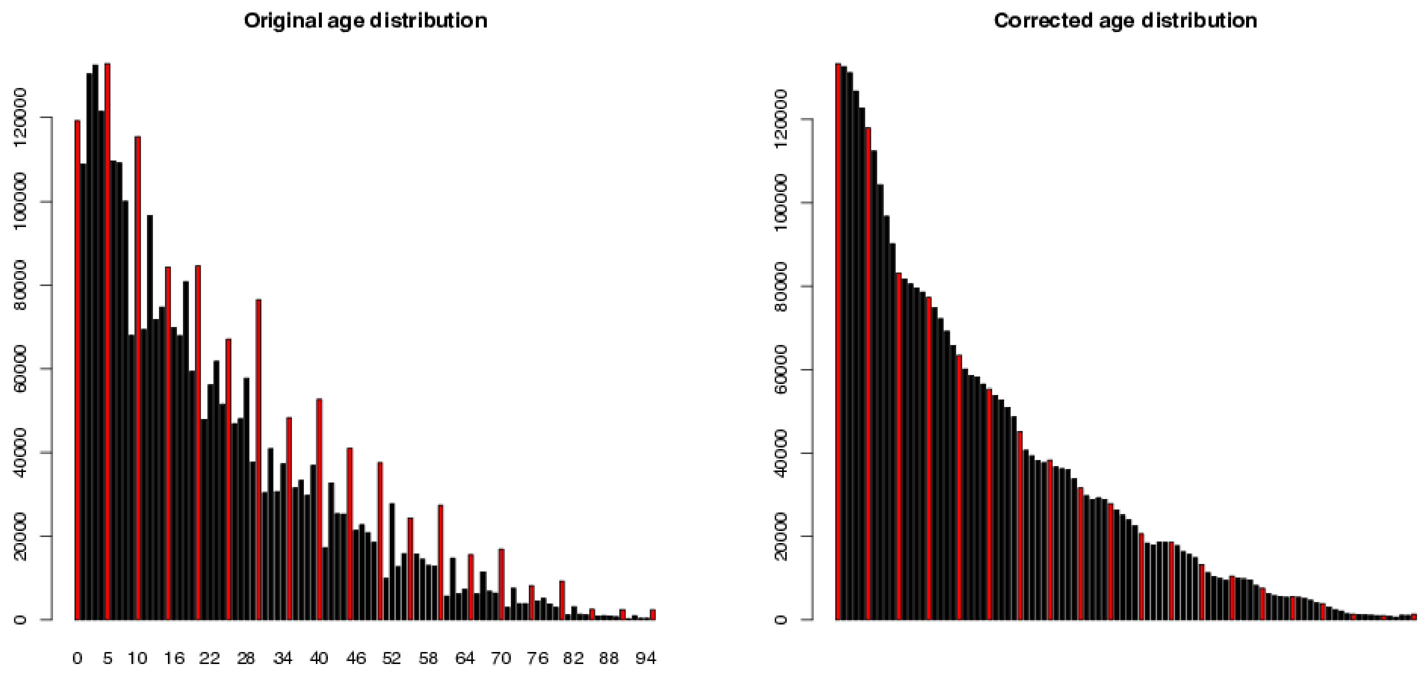
Results

```
spCdfplot(synthPadj, "netIncome", cond="db040", layout=c(3, 3))
```



Other feature of simPop - age heaping

Correct for age heaping using truncated (log-)normal distributions on individual level (function **correctHeap()**)



Conclusions

- Structure of original input data is preserved
- Margins of synthetic populations are calibrated
- The synthetic populations are confidential
- Code of **simPop** is quite efficient
- Many methods are ready to be used