Synthetic Data: Modelling and Generating Complex Close-to-Reality Data for Public Use

Matthias Templ,
Bernhard Meindl
Statistics Austria &
Vienna Uni. of Techn.
Mai 2015

ODAM 2015, Olomouc, CZ

Where to need synthetic populations

Population data are needed

- designing better social and economic policies and programs through **micro-simulation** and **agent-based modeling**
- to compare methods in **design-based simulation studies**
- **teaching**, especially teaching of survey methods
- public- and scientific-use files including no risk of disclosure

Note: simulation of survey samples $\subset$ simulation of populations

History and fields of applications

- Clarke et al. (1984): created a initial population from aggregated data for British health district authorities.
- estimation of the demand of water by Clarke and Holm (1987), Williamson et al. (1998)
- from 1998 onwards the research field is considerable growing, not to say dramatically grown.
- examples: health planning (Brown and Harding 2002, Tomintz et al. 2008, Smith et al. 2011), transportation (Beckman et al. 1996, Barthelemy and Toint 2013) or environmental planning (Williamson et al. 2002), and it could be applied to simulate deseases, climate change, demographic change and economic change.

History and fields of applications

- Clarke et al. (1984): created a initial population from aggregated data for British health district authorities.
- estimation of the demand of water by Clarke and Holm (1987), Williamson et al. (1998)
- from 1998 onwards the research field is considerable growing, not to say dramatically grown.
- examples: health planning (Brown and Harding 2002, Tomintz et al. 2008, Smith et al. 2011), transportation (Beckman et al. 1996, Barthelemy and Toint 2013) or environmental planning (Williamson et al. 2002), and it could be applied to simulate deseases, climate change, demographic change and economic change.

History and fields of applications

- Clarke et al. (1984): created a initial population from aggregated data for British health district authorities.
- estimation of the demand of water by Clarke and Holm (1987), Williamson et al. (1998)
- from 1998 onwards the research field is considerable growing, not to say dramatically grown.
- examples: health planning (Brown and Harding 2002, Tomintz et al. 2008, Smith et al. 2011), transportation (Beckman et al. 1996, Barthelemy and Toint 2013) or environmental planning (Williamson et al. 2002), and it could be applied to simulate deseases, climate change, demographic change and economic change.

Close-to-reality data

Conditions:

- actual sizes of regions and strata need to be reflected;
- marginal distributions and interactions between variables should be represented correctly;
- hierarchical and cluster structures has 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.

Close-to-reality data

Conditions:

- actual sizes of regions and strata need to be reflected;
- marginal distributions and interactions between variables should be represented correctly;
- hierarchical and cluster structures has 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.

Close-to-reality data

Conditions:

- actual sizes of regions and strata need to be reflected;
- marginal distributions and interactions between variables should be represented correctly;
- hierarchical and cluster structures has 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.

Close-to-reality data

Conditions:

- actual sizes of regions and strata need to be reflected;
- marginal distributions and interactions between variables should be represented correctly;
- hierarchical and cluster structures has 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.

Close-to-reality data

Conditions:

- actual sizes of regions and strata need to be reflected;
- marginal distributions and interactions between variables should be represented correctly;
- hierarchical and cluster structures has 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.

Dependency on method choice

the choice of methods often depends on the available information

- only aggregated information from samples
- survey samples
- known marginal distributions of the population
- multiple data sources

e.g, no sample data $\rightarrow$ simulate from multiple tables by drawing conditionally from marginal distributions.

Dependency on method choice

the choice of methods often depends on the available information

- only aggregated information from samples
- survey samples
- known marginal distributions of the population
- multiple data sources

e.g, no sample data $\rightarrow$ simulate from multiple tables by drawing conditionally from marginal distributions.

Calibration of sample data

- Sample weights are calibrated to fit exactly known marginal population totals.
- $S_i = 1$ if individual i is sampled, $S_i = 0$ otherwise.
  Problem: estimate $Y = \sum_{i=1}^{N}$ for a population of size N.
  Horwitz-Thomposon estimator

$$\widehat{Y}_d = \sum_{i:S_i=1} d_i y_i \quad , \tag{1}$$

  with $d_i = 1/\pi_i$ the inverse of the first order inclusion probablity of individual i in the population.
- If an auxiliary varible x is available from the sample with the condition that the population total $X = \sum_{i=1}^{N} x_i$ is known, usually $\sum_{i:S_i=1} d_i x_i \neq X$. The aim is to find new (calibrated) weights w_i with $\widehat{Y}_w = \sum_{i:S_i=1} w_i y_i$ where $\sum_{i:S_i=1} w_i x_i = X$ and $\sum_{i:S_i=1} w_i = N$.

Calibration: IPF

Iterative Proportional Fitting / Raking to

- estimate joint-distribution of the true population given the information from the sample for a number of control variables in a way that it is consistent with known totals (marginal distributions).
- iterative procedures or regression-based approaches to reach this goal.
- Iterative proportional updating (IPU) controls for individual- and household-level control variables at the same time by solving a mathematical optimization problem heuristically.

Methods to simulate synthetic populations

three kinds of methods

1. **synthetic reconstruction / deterministic reweighting** methods to simulate basic populations (based on conditional probabilities). Used in combination with calibration methods (IPF, IPU, HIPF) to calibrate samples
2. **combinatorical optimization** techniques to achieve known charateristics for populations and to combine synthetic population with detailed geographical information - SA, GA
3. **model-based methods** to simulate synthetic close-to-reality populations using regression techniques

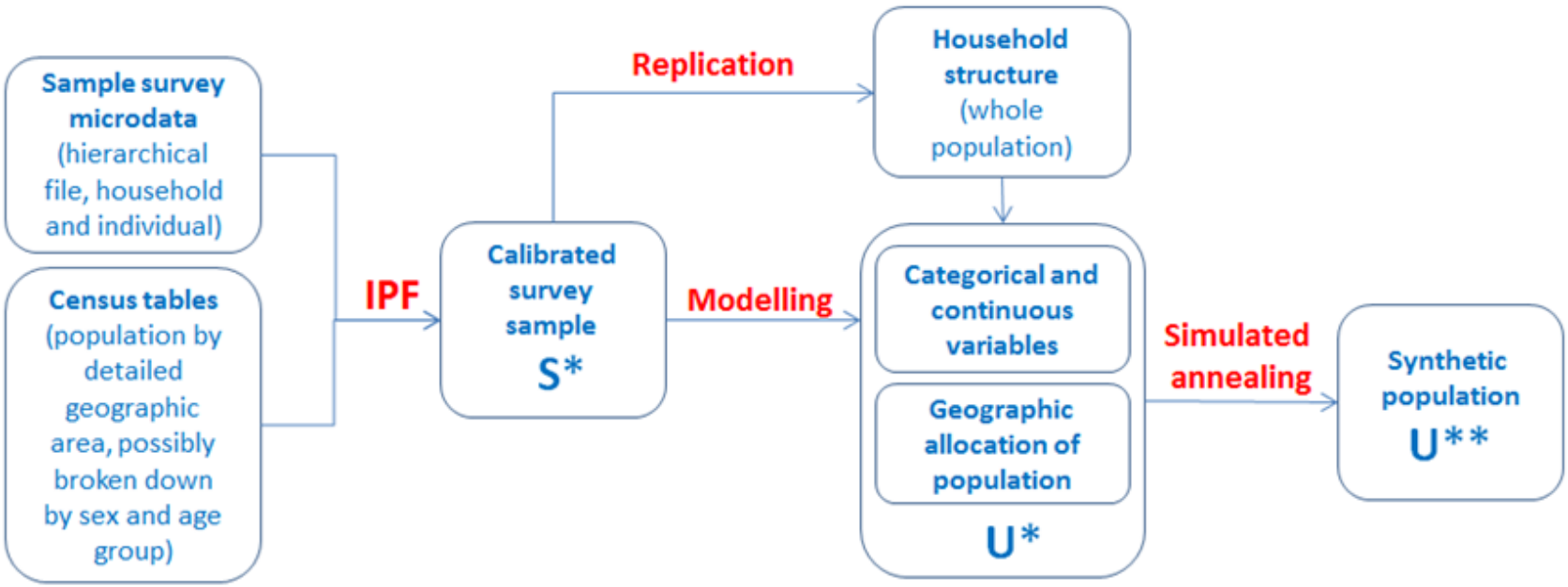
Synthetic reconstruction

Synthetic reconstruction techniques are the most frequently used methods.

- combining information from two sources of data, aggregated data in the form of census tables and survey micro-dataset representative of the population of interest (the seed)
- **1) Estimation:** a joint distribution is estimated using both sources of data.
- **2) Selection:** Individuals are randomly selected from the seed dataset and added to the synthetic population so that the joint probabilities calculated in the previous step are respected.

Complex to generate realistic household structures and problems with zero-margins $\rightarrow$ we propose to use model-based methods instead.

Model-based approach



Model-based simulation

In general, the procedure consists of four steps (Alfons et al. 2011):

- The setup of the household structure;
- the simulation of categorical variables;
- the simulation of continuous variables;
- (the splitting continuous variables into components.)

Stratification allows to account for heterogenities such as regional differences. **Sampling weights** are considered in each step to ensure high similarity of expected and realized values.

Simulation of the basic structure

- The household structure is simulated separately for each combination of strata and household size.
- The number of households is estimated using the Horvitz-Thompson estimator.
- As few variables as possible (due to confidentiality reasons) are simulated using Alias sampling.
- This builds up a realistic structure of the few basic variables chosen.

Additional variables are then simulated using a regression-based approach.

Simulation of the basic structure

- The household structure is simulated separately for each combination of strata and household size.
- The number of households is estimated using the Horvitz-Thompson estimator.
- As few variables as possible (due to confidentiality reasons) are simulated using Alias sampling.
- This builds up a realistic structure of the few basic variables chosen.

Additional variables are then simulated using a regression-based approach.

Additional variables: fitting on sample

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

$\longrightarrow$ built model-matrix from the predictors to predict $\boldsymbol{x}_{j+1}$

$\longrightarrow$ fit of β's (multinomial regression, naivebayes, two-step approaches, ...)

Additional variables: prediction on the population

$$\textit{population} \quad \boldsymbol{U} = \left(\overbrace{\begin{array}{cccc} \widehat{x}_{1,1} & \widehat{x}_{1,2} & \cdots & \widehat{x}_{1,j} \\ \widehat{x}_{2,1} & \widehat{x}_{2,2} & \cdots & \widehat{x}_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{x}_{N,1} & \widehat{x}_{n,2} & \cdots & \widehat{x}_{N,j} \end{array}}^{\widehat{\beta} \times \text{ pred. } \approx} \overbrace{\begin{array}{c} \widehat{x}_{1,j+1} \\ \widehat{x}_{1,j+1} \\ \vdots \\ \vdots \\ \widehat{x}_{1,j+1} \end{array}}^{\widehat{x}_{j+1}} \right)$$

Generally, not the expected values are used.

Simulation of categorical variables

- **regression** methods are applied on the sample data (multinomial, naivebayes or synthetic reconstruction)
- The parameters (regression coefficients) obtained from the model fit **on the sample** are used to simulate the variable of interest on **population level**.
- This is based on draws with estimated probabilities for each category on individual basis.

Simulation of continuous variables

- almost the same approach as before, but either a
  - a multinomial model with random draws from (previously builded) resulting **categories** or
  - a **two-step regression model** with random error terms is used to simulate the new variable

Random errors (noise) by drawing from the residuals are added.

Simulation of continuous variables

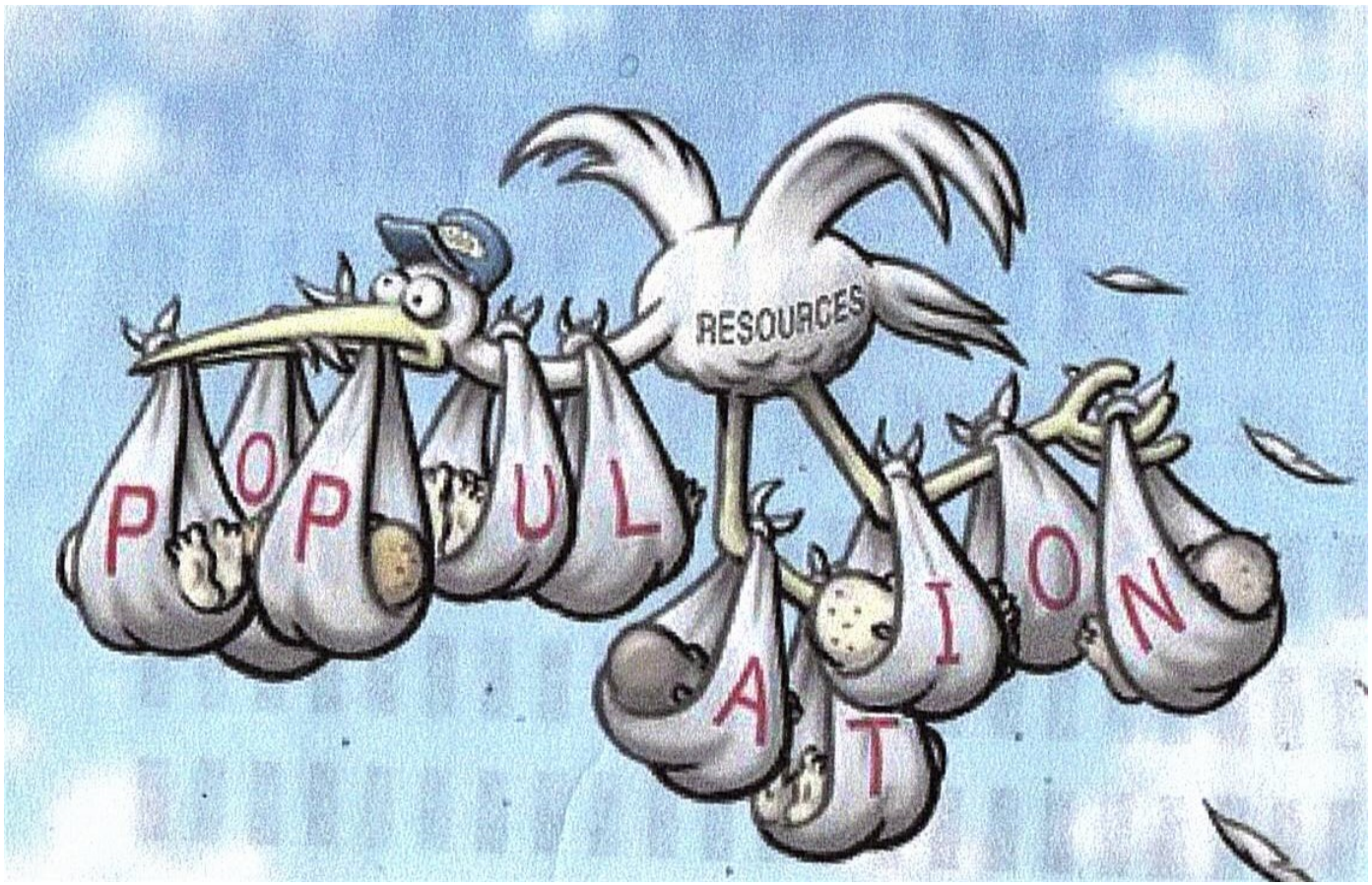
- almost the same approach as before, but either a
  - a multinomial model with random draws from (previously builded) resulting **categories** or
  - a **two-step regression model** with random error terms is used to simulate the new variable

Random errors (noise) by drawing from the residuals are added.

Adjusting the population

- Estimation of means, totals, etc. **is unbiased** for synthetic populations, but a **single realization** of a population may differ in some known population characteristics.
- Combinatorial optimization methods are used to calibrate to known population characteristics.
- Basically a larger population is generated and households are assigned with 0/1 weights
- Exchange 0/1 weights in a clever way as long as known population characteristics are estimated precisely enough.

Going to applications simulating the Austrian population



R package simPop

- Developed with support of the International Household Survey Network, DFID Trust Fund TF011722
- contains all mentioned methods (and more)
- highly object-oriented approach
- let you produce synthetic confidential data
- efficiently programmed to work for (very) large data sets
- parallel computing is automatically be applied
- now used on large scale by the world bank

Example: EU-SILC

```
require(simPop)
str(origData)
```

```
## 'data.frame': 11725 obs. of  18 variables:
##  $ db030    : int  1 1 2 3 4 4 4 5 5 5 ...
##  $ hsize    : int  2 2 1 1 3 3 3 5 5 5 ...
##  $ db040    : Factor w/ 9 levels "Burgenland","Carinthia",..: 4 4 7 5 7 7 7 4 4 4 ...
##  $ age      : int  72 66 56 67 70 46 37 41 35 9 ...
##  $ rb090    : Factor w/ 2 levels "male","female": 1 2 2 2 2 1 1 1 2 2 ...
##  $ pl030    : Factor w/ 7 levels "1","2","3","4",..: 5 5 2 5 5 3 1 1 3 NA ...
##  $ pb220a   : Factor w/ 3 levels "AT","EU","Other": 1 1 1 1 1 1 3 1 1 NA ...
##  $ netIncome: num  22675 16999 19274 13319 14366 ...
##  $ py010n   : num  0 0 19274 0 0 ...
##  $ py050n   : num  0 0 0 0 0 ...
##  $ py090n   : num  0 0 0 0 0 ...
##  $ py100n   : num  22675 0 0 13319 14366 ...
##  $ py110n   : num  0 0 0 0 0 0 0 0 0 NA ...
##  $ py120n   : num  0 0 0 0 0 0 0 0 0 NA ...
##  $ py130n   : num  0 16999 0 0 0 ...
##  $ py140n   : num  0 0 0 0 0 0 0 0 0 NA ...
##  $ db090    : num  7.82 7.82 8.79 8.11 7.51 ...
##  $ rb050    : num  782 782 879 811 751 ...
```

Structure your data (once to be defined)

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

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

```
inp
```

```
## --------------
## survey sample of size 11725 x 20
##
## Selected important variables:
##
## household ID: db030
## personal ID: pid
## variable household size: hhsize
## sampling weight: rb050
## strata: db040
## --------------
```

Additional information

(external) Population charateristics on EU-SILC variables as data frame or n-dimensional table (here a 2-dimensional table):

```
totalsRGtab
```

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

Calibration to this given known totals:

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

Simulating the basic structure

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

```
class(synthP)

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

The resulting output object ("synthP") is of class *simPopObj*. As already mentioned, various functions can be directly applied to objects of that class.

Simulation of categorical variables

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

## dealing with level pl030
## dealing with level pb220a
```

```
synthP

##
## --------------
## synthetic population  of size
##  8504755 x 10
##
## build from a sample of size
## 11725 x 19
## --------------
##
## variables in the population:
## db030,hsize,db040,age,rb090,db040.1,pid,weight,pl030,pb220a
```

Simulating continuous variables

```
# multinomial model with random draws
synthP <- simContinuous(synthP, additional = "netIncome",
    upper = 2e+05, equidist = FALSE)
```

To simulate components use `simComponents()`, to simulate finer regional variables (like districts), use `simInitSpatial()`.

Census information to calibrate

assume you have (again) external information (n-dimensional table), here e.g. marginals on *region* $\times$ *gender* $\times$ *exonomic status*.
We add these marginals to the object and calibrate afterwards

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

```
# calibration by 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=FALSE)
```

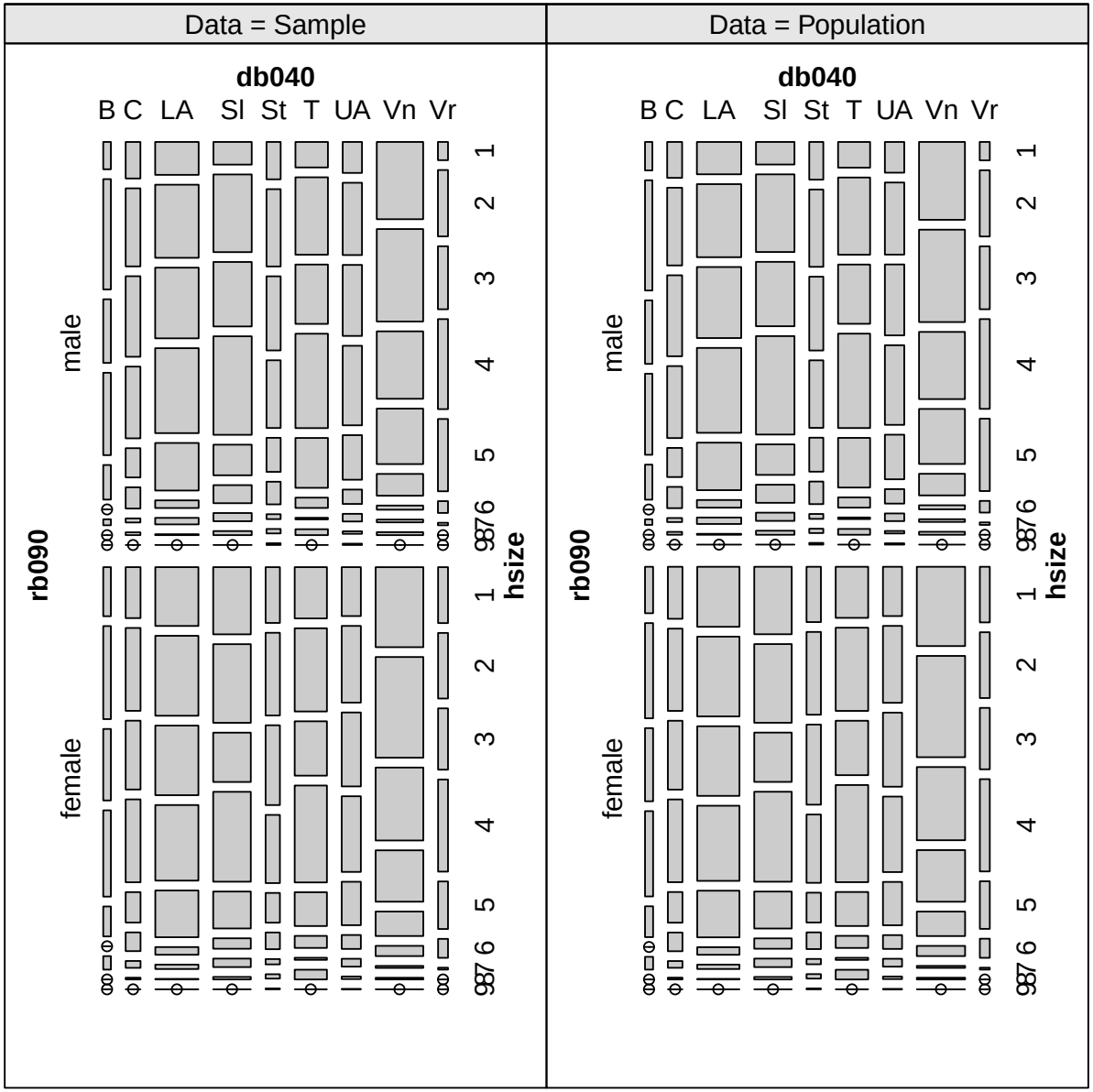
To speed up the computations, parallel computing is applied automatically.

Continued, we end up with

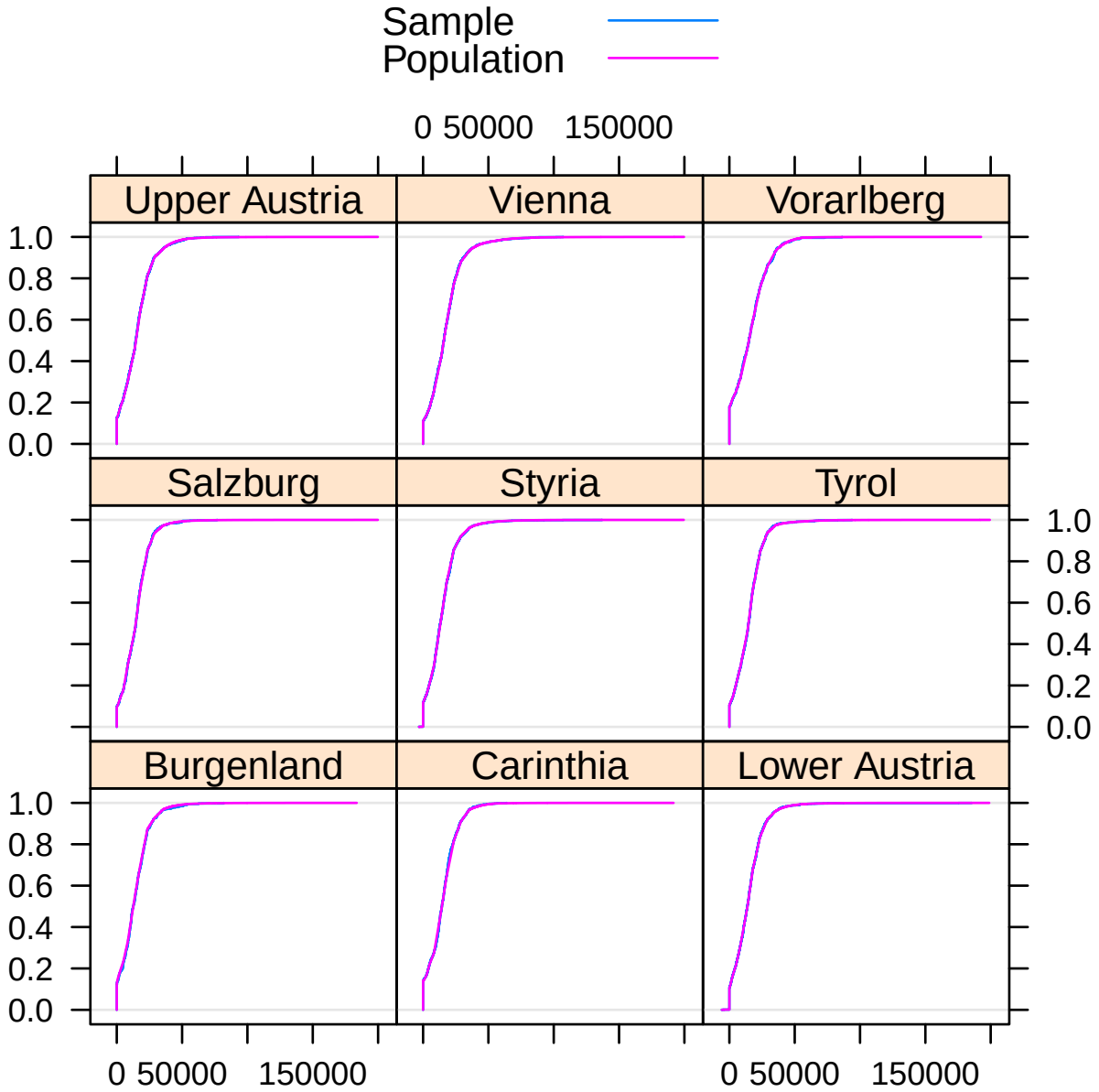
```
synthP

##
## --------------
## synthetic population  of size
##  8504755 x 20
##
## build from a sample of size
## 11725 x 20
## --------------
##
## variables in the population:
## db030,hsize,db040,age,rb090,db040.1,pid,weight,pl030,
## pb220a,netIncomeCat,netIncome,py010n,py050n,py090n,
## py100n,py110n,py120n,py130n,py140n
```

Results


Data = Sample
Data = Population
db040
B C LA Sl St T UA Vn Vr
rb090
male
female
hsize
1 2 3 4 5 6 7 8 9

Results


Sample
Population
Upper Austria
Vienna
Vorarlberg
Salzburg
Styria
Tyrol
Burgenland
Carinthia
Lower Austria
0 50000 150000
1.0
0.8
0.6
0.4
0.2
0.0

Conclusions

- margins of synthetic populations are calibrated
- all statistics can be very precisely estimated
- the synthetic populations are confidential
- code of **simPop** is highly efficient
- many other methods are included
- large applications on data from world bank follow

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, 2011. URL http://dx.doi.org/10.1007/s10260-011-0163-2.

J. Barthelemy and P.L. Toint. Synthetic population generation without a sample. Transportation Science, 47(2):266–279, 2013. URL http://dblp.uni-trier.de/db/journals/transci/transci47.html#BarthelemyT13.

R.J. Beckman, K.A. Baggerly, and M.D. McKay. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice, 30(6):415–429, 1996. URL http://EconPapers.repec.org/RePEc:eee:transa:v:30:y:1996:i:6:p:415-429.

L. Brown and A. Harding. Social Modelling and Public Policy: Application of Microsimulation Modelling in Australia. Journal of Artificial Societies and Social Simulation, 5(4):6, 2002. URL http://ideas.repec.org/a/jas/jasssj/2002-33-1.html.

G.P. Clarke, C. Clarke, M. Birkin, P.H. Rees, and A.G. Wilson. A Strategic Planning Simulation Model of a District Health Service System: The In-patient Component and Results. Number Bd. 385-389 in A Strategic Planning Simulation Model of a District Health Service System: The In-patient Component and Results. School of Geography, University of Leeds, 1984.

M. Clarke and E. Holm. Microsimulation methods in spatial analysis and planning. Arbetsrapport från CERUM. Umeå universitet. Geografiska inst., 1987. URL http://books.google.at/books?id=fpR2RAAACAAJ.

D.M. Smith, J.R. Pearce, and K. Harland. Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? an example of smoking prevalence in new zealand. Health Place, 17(2):618–24, 2011. URL http://www.biomedsearch.com/nih/Can-deterministic-spatial-microsimulation-model/21257335.html.

Melanie N. Tomintz, Graham P. Clarke, and Janette E. Rigby. The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services. Area, 40(3):341–353, 2008. doi: 10.1111/j.1475-4762.2008.00837.x. URL http://dx.doi.org/10.1111/j.1475-4762.2008.00837.x.

P. Williamson, M. Birkin, and P.H. Rees. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environ Plan A, 30(5):785–816, 1998.

P. Williamson, G. Mitchell, and A. T. McDonald. Domestic water demand forecasting: A static microsimulation approach. Water and Environment Journal, 16(4):243–248, 2002. URL http://dx.doi.org/10.1111/j.1747-6593.2002.tb00410.x.