Matthias Templl,
Bernhard Meindl
Statistics Austria &
Vienna Uni. of Techn. &
data-analysis OG &
World Bank &
Sept 17, 2014

Eivissa, Spain

Methods and Tools for the Generation of Synthetic Populations

STATISTIK AUSTRIA
Die Informationsmanager

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

STATISTIK AUSTRIA
Die Informationsmanager

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

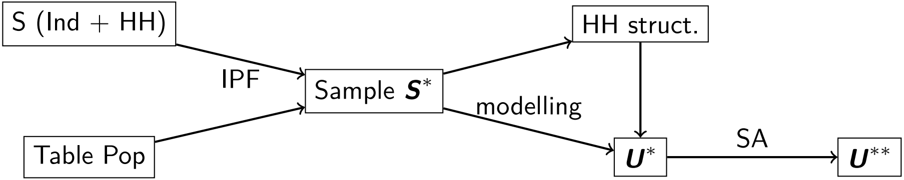
Conditions:

- 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.

Comparison of methods

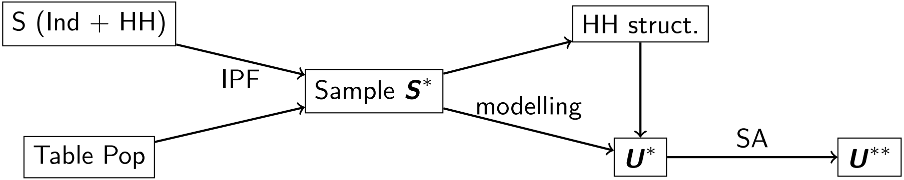
advantages	disadvantages
synthetic reconstruction (and IPU, IPF, HIPF)	
▸ combination of different data sources is possible	▸ continuous data cannot be simulated ▸ complicated to consider the household structure ▸ zero-cells and zero-marginals
modelling & prediction	
▸ the correlation structure is preserved	▸ computation complex for simulating multinomial variables ▸ results depend on the order of variables to be simulated
combinatorial optimization	
▸ data structure is preserved	▸ high computational costs
super-population models (Copulas)	
▸ computationally fast	▸ not possible to simulate complex microdata structures

Model-based approach



... in reallity it is often even more complex.

Model-based approach



... in reallity it is often even more complex.

Model-based simulation

In general, the procedure consists of four steps:

- 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.

Simulation of categorical variables

- **multinomial logistic regression** is applied on the sample data, having the variable to simulate as response and chosen variables **of the sample** as predictors.
- The predictor variables are those variables **from the sample** that are **already simulated for the population**.
- The parameters (regression coefficients) obtained from the model fit **on the sample** are then used to calculate the variable of interest on **population level** by a linear combination (determined by the estimated regression coefficients) of the already simulated variables.

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 error (noise) by drawing from the residuals need to be 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 error (noise) by drawing from the residuals need to be added.

R package synthPop

- Developed with support of the International Household Survey Network, DFID Trust Fund TF011722 and data-analysis OG
- contains all mentioned methods
- highly object-oriented approach, similar to sdcMicro
- let you produce synthetic confidential data
- efficiently programmed to work for (very) large data sets
- parallel computing is automatically be applied

Example: EU-SILC

```
require(synthPop)
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)
```

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

```
class(synthP)

## [1] "synthPopObj"
## attr(,"package")
## [1] "synthPop"
```

The resulting output object ("synthP") is of class *synthPopObj*. 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 (next slide)

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

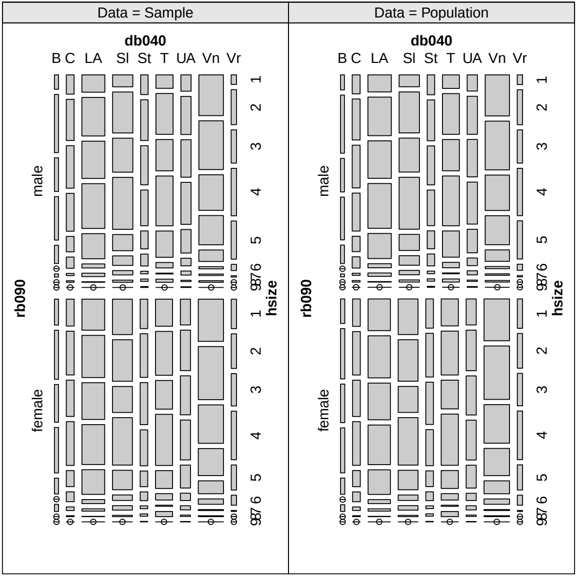
```
# calibration by simulated annealing
synthPadj <- calibPop(synthP, split="db040", temp=1,
      eps.factor=0.00005, maxiter=200, temp.cooldown=0.975,
      factor.cooldown=0.85, min.temp=0.001, verbose=FALSE)
```

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

Continued, we end up with

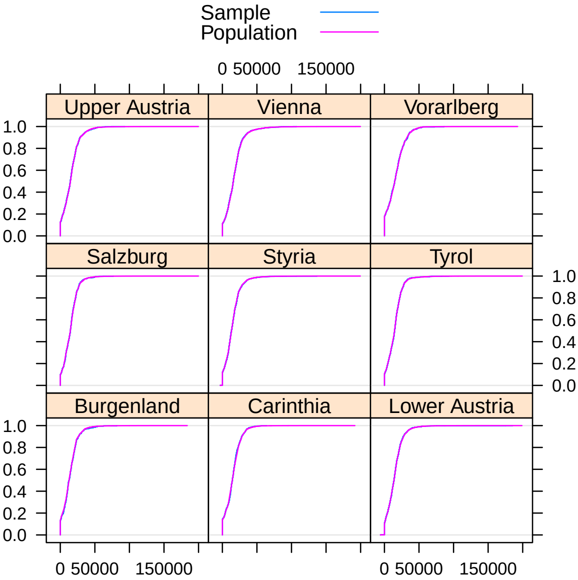
```
synthP

##
## --------------
## 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
0 50000 150000
Upper Austria
Vienna
Vorarlberg
Salzburg
Styria
Tyrol
Burgenland
Carinthia
Lower Austria
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 **synthPop** is highly efficient
- many other methods are included
- large applications on data from world bank follow