5/24/2017, SurveyConf 2017

## Submit your work



## Synthetic data?

* Ralf mentioned in 2006 that it is important and we continued some of his previous work in the AMELI project.
* Clearly, synthetic data generated, e.g., with

mvtnorm::rmvnorm(n = 500,   
 mean = c(1,2),   
 sigma = matrix(c(4,2,2,3), ncol=2))

X <- T %\*% t(B) + E # component model

model-based imputation methods suggested by Rubin (1993) and many other authors or using Copulas, machine and deep learning methods –> too simplistic

## 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

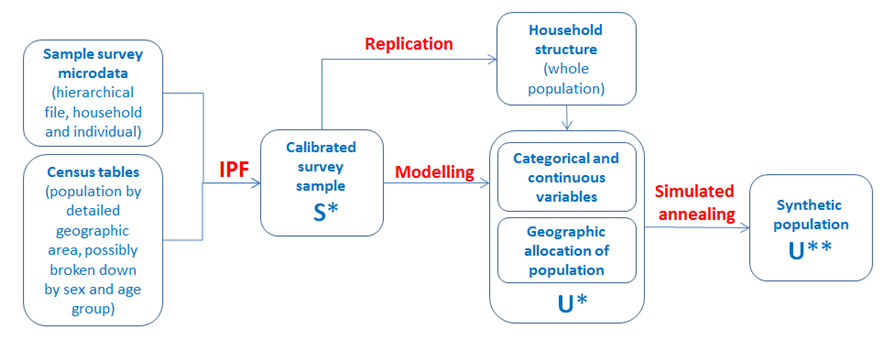
* choice of methods depends on available information:
  + census
  + survey samples
  + aggregated information from samples
  + known marginal distributions from population

## Methods

1. **Synthetic reconstruction** / **Deterministic reweighting** are based on conditional probabilites. Often used in combination with calibration methods (IPF, IPU, HIPF)
2. **Combinatorial optimization** are used to calibrate populations and to enrich a population e.g. with detailed geographical information (SA, GA)
3. **Model-based methods** for the simulation of close-to-reality  
     
   populationens using regression methods

## Model-based approach - workflow

Example workflow individuals within households survey data:



## 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

* Household structure (core-variables): simulated separately for each combination of strata and household size.
* Number of households: estimated using the HT-estimator
* As few variables as possible (due to confidentiality reasons) are simulated using a sampling approach
* This builds up a realistic structure of the core variables
* Finally, additional variables are simulated using a regression/ML-based approach using either
* all existing variables
* a defined model
* basic structural variables

## (Simplified) Rule

* fit on the sample data
* using this information, predict on population data

E.g. citizenship is available in the sample but not in the population. Build a model on the **sample data** using predictors that are available on the sample and population. Predict citizenship on the **population data**.

## 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)

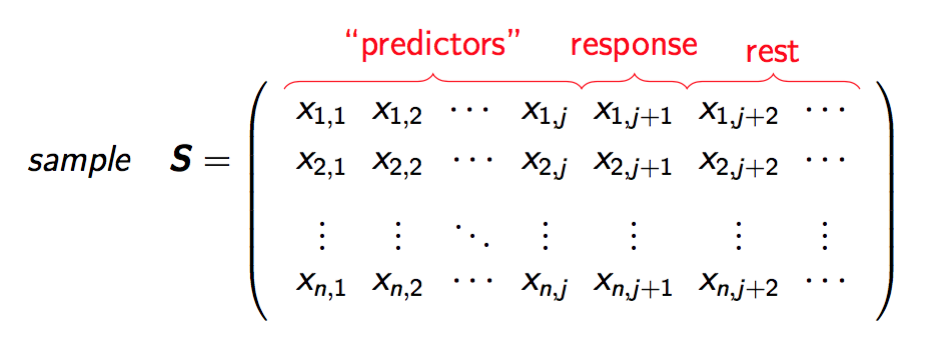
We will see later in a demo:

#### **simStructure()**

## Model-based approach

* Input: survey data
* Model: variable ∼ covariates (better: model matrix) fitted on sample survey. First predictors are the basic structure variables
* Regression coefficients used to predict on population (do not use expected values)

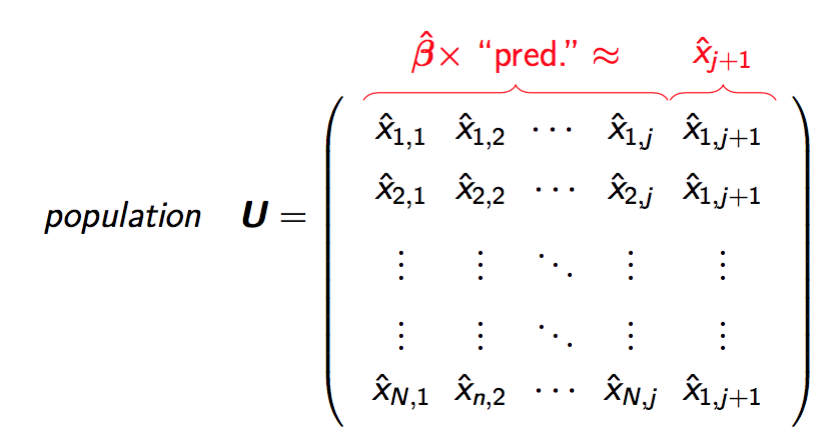
## Model-based approach - fitting



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

\(\longrightarrow\) estimation of the \(\boldsymbol{\beta}\)'s

## Model-based approach - prediction



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

## Model-based approach - categorical variables

Estimation of the \(\boldsymbol{\beta}\)'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
* **cforest**: 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

#### **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()**)

## Applications on real-world data

* AMELI EU-FP7 project
* Templ and Alfons (2010) (EU-SILC)
* Alfons, Kraft, Templ, and Filzmoser (2011) (Risk)
* Templ and Filzmoser (2014) (Employer-Employee)
* EU project SGA on PUF (EU-SILC for many countries)
* Frazier and Alfons (2017) (Ghana)
* Templ, Kowarik, and Meindl (2016) (EU-SILC)
* World bank project to simulate a Census for Mauritania

## R package simPop

* Templ, Kowarik, and Meindl (2016), Journal of Statistical Software (accepted)
* latest version on [CRAN](http://cran.r-project.org/web/packages/simPop/index.html)
* Demo: we produce synthetic confidential data (simplified!)
* parallel computing is applied automatically
* efficient implementation

## Special classes in simPop

simPop uses S4 classes, the most important classes defined:

* **dataObj**: Contains **information** on the population and **survey data** to be used as input for the generation of the synthetic population. Typically, information on the variables containing the household and person IDs, household size, sampling weights, stratification information, and type of data (i.e., sample or a population).
* **simPopObj**: Contains information on the **sample** (in slot sample), the **population** (slot pop), and optionally some margins in the form of a **table** (slot table). Objects in slot sample and pop must be objects of class **dataObj**.

## Example: EU-SILC

* EU-SILC is one of the most popular data sets in Europe.
* It is used to measure social cohesion and poverty in Europe

library("simPop")   
data("eusilcS")   
origData <- eusilcS  
origData$rb050 <- origData$rb050 \* 100

library("simPop")  
str(origData[,1:7])

## 'data.frame': 11725 obs. of 7 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 ...

str(origData[,8:18])

## 'data.frame': 11725 obs. of 11 variables:  
## $ 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 ...

## 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"

## Print the dataObj

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  
## --

## 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

## Simulating continuous variables

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

##   
## --  
## synthetic population of size   
## 8182010 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,netIncome

* Simulation of components with **simComponents()**

## Census information to calibrate

* Assumption: external information (n-dimensional table) is available, e.g marginals on *region* \(\times\) *gender* \(\times\) *exonomic status*.

margins

## db040 rb090 pb220a freq  
## 1 Burgenland male AT 797  
## 2 Lower Austria male AT 4363  
## 3 Vienna male AT 3749  
## 4 Carinthia male AT 1568  
## 5 Styria male AT 3139  
## 6 Upper Austria male AT 3741  
## 7 Salzburg male AT 1379  
## 8 Tyrol male AT 1566  
## 9 Vorarlberg male AT 819  
## 10 Burgenland female AT 821  
## 11 Lower Austria female AT 4426  
## 12 Vienna female AT 4357  
## 13 Carinthia female AT 1729  
## 14 Styria female AT 3354  
## 15 Upper Austria female AT 3962  
## 16 Salzburg female AT 1508  
## 17 Tyrol female AT 1886  
## 18 Vorarlberg female AT 902  
## 19 Burgenland male EU 26  
## 20 Lower Austria male EU 53  
## 21 Vienna male EU 201  
## 22 Carinthia male EU 34  
## 23 Styria male EU 29  
## 24 Upper Austria male EU 61  
## 25 Salzburg male EU 27  
## 26 Tyrol male EU 80  
## 27 Vorarlberg male EU 21  
## 28 Burgenland female EU 29  
## 29 Lower Austria female EU 48  
## 30 Vienna female EU 307  
## 31 Carinthia female EU 50  
## 32 Styria female EU 50  
## 33 Upper Austria female EU 87  
## 34 Salzburg female EU 57  
## 35 Tyrol female EU 82  
## 36 Vorarlberg female EU 15  
## 37 Burgenland male Other 10  
## 38 Lower Austria male Other 222  
## 39 Vienna male Other 622  
## 40 Carinthia male Other 32  
## 41 Styria male Other 119  
## 42 Upper Austria male Other 270  
## 43 Salzburg male Other 195  
## 44 Tyrol male Other 102  
## 45 Vorarlberg male Other 94  
## 46 Burgenland female Other 7  
## 47 Lower Austria female Other 201  
## 48 Vienna female Other 465  
## 49 Carinthia female Other 37  
## 50 Styria female Other 112  
## 51 Upper Austria female Other 227  
## 52 Salzburg female Other 215  
## 53 Tyrol female Other 138  
## 54 Vorarlberg female Other 94

## 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, calculation time between 45 min and some hours)

## Quality and disclosure risk

Comparison of sample information (and known census information) and synthetic population.

Utility measures varies from point and variance estimation of indicators, visual comparisons, model results.

Disclosure risk: Templ and Alfons (2010)

Utility:

\* quality indicators (Templ 2015, 2017)

\* SILC: Alfons, Kraft, Templ, Filzmoser (2011b), Bergeat et al. (2016)

\* Employer-Employee data: Templ and Filzmoser (2014)

## Results

tab <- spTable(synthP, select=c("rb090","db040","hsize"))  
spMosaic(tab, labeling=labeling\_border(abbreviate=c(db040=TRUE)))

## Results

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

## Results

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

## Other features of simPop

age heaping

## Other features of simPop

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

Acknowledgement: World bank for several funded projects.

Main programming:

[http://www.data.analysis.at](http://www.data-analysis.at)

## References

[1] A. Alfons, S. Kraft, M. Templ, et al. "Simulation of close-to-reality population data for household surveys with application to EU-SILC". In: *Statistical Methods & Applications* 20.3 (2011), pp. 383-407. <URL: <http://dx.doi.org/10.1007/s10260-011-0163-2>>.

[2] J. Barthelemy and P. Toint. "Synthetic Population Generation Without a Sample." In: *Transportation Science* 47.2 (2013), pp. 266-279. <URL: <http://dblp.uni-trier.de/db/journals/transci/transci47.html#BarthelemyT13>>.

[3] R. Beckman, K. Baggerly and M. McKay. "Creating synthetic baseline populations". In: *Transportation Research Part A: Policy and Practice* 30.6 (1996), pp. 415-429. <URL: [http://EconPapers.repec.org/RePEc:eee:transa:v:30:y:1996:i:6:p:415-429](http://econpapers.repec.org/RePEc:eee:transa:v:30:y:1996:i:6:p:415-429)>.

[4] L. Brown and A. Harding. "Social Modelling and Public Policy: Application of Microsimulation Modelling in Australia". In: *Journal of Artificial Societies and Social Simulation* 5.4 (2002), p. 6. <URL: <http://ideas.repec.org/a/jas/jasssj/2002-33-1.html>>.

[5] G. Clarke, C. Clarke, M. Birkin, et al. *A Strategic Planning Simulation Model of a District Health Service System: The In-patient Component and Results*. A Strategic Planning Simulation Model of a District Health Service System: The In-patient Component and Results Bd. 385-389. School of Geography, University of Leeds, 1984.

[6] M. Clarke and E. Holm. *Microsimulation methods in spatial analysis and planning*. Arbetsrapport frCERUM. Umeuniversitet. Geografiska inst., 1987. <URL: <http://books.google.at/books?id=fpR2RAAACAAJ>>.

[7] T. Frazier and A. Alfons. *Generating a close-to-reality synthetic population of Ghana*. 2016.

[8] D. Smith, J. 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." In: *Health Place* 17.2 (2011), pp. 618-24. <URL: <http://www.biomedsearch.com/nih/Can-deterministic-spatial-microsimulation-model/21257335.html>>.

[9] M. Templ and A. Alfons. "Disclosure risk of synthetic population data with application in the case of EU-SILC". In: *Privacy in Statistical Databases*. Ed. by J. Domingo-Ferrer and E. Magkos. Vol. 6344. Lecture Notes in Computer Science. Heidelberg: Springer, 2010, pp. 174-186.

[10] M. Templ and P. Filzmoser. "Simulation and quality of a synthetic close-to-reality employer-employee population". In: *Journal of Applied Statistics* 41.5 (2014), pp. 1053-1072.

[11] M. Templ, A. Kowarik and B. Meindl. "Simulation of Synthetic Complex Data: The R-Package simPop". In: *Journal of Statistical Software* (2016). accepted for publication, pp. 1-39.

[12] M. N. Tomintz, G. P. Clarke and J. E. Rigby. "The geography of smoking in Leeds: estimating individual smoking rates and the implications for the location of stop smoking services". In: *Area* 40.3 (2008), pp. 341-353. DOI: 10.1111/j.1475-4762.2008.00837.x. <URL: <http://dx.doi.org/10.1111/j.1475-4762.2008.00837.x>>.

[13] P. Williamson, M. Birkin and P. Rees. "The estimation of population microdata by using data from small area statistics and samples of anonymised records". In: *Environ Plan A* 30.5 (1998), pp. 785-816.

[14] T. Williamson. *Knowledge and its Limits*. Oct. 10, 2002. DOI: 10.1093/019925656x.001.0001. <URL: <http://dx.doi.org/10.1093/019925656x.001.0001>>.