Matthias Templ
QM @ Statistik Austria
März 2016

Mittwochseminar

Marry me!
Offizielle Statistik  Forschung mittels synthetischer Daten
(& Methoden der synthetischen Datengenerierung)

Danksagung

Kooperationen:

- Software **simPopulation** (Alfons, Kraft, Templ, and Filzmoser 2011)
- Software **simPop** (Templ, Kowarik, and Meindl 2016a)

Überblick

STATISTIK AUSTRIA
Die Informationsmanager

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

Vorurteile

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

- Wir haben doch Populationen im Haus!
- Andere arbeiten auch nicht mit synth. Daten!
- Synthetische Daten sind keine Echtdaten!
- Mit synth. Daten verlieren wir Glaubwürdigkeit!
- Wir haben wichtigeres zu tun!
- Steckenpferd der Wissenschaft!

Vorurteile

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

- Wir haben doch Populationen im Haus!
- Andere arbeiten auch nicht mit synth. Daten!
- Synthetische Daten sind keine Echtdaten!
- Mit synth. Daten verlieren wir Glaubwürdigkeit!
- Wir haben wichtigeres zu tun!
- Steckenpferd der Wissenschaft!

Synthetische Populationen. Warum?

1) Viele interessante Variablen: nur durch Surveys
   - Anreicherung von Populationen mit Survey-Information
2) Simulationsstudien zur Evaluierung und Entwicklung von Methoden
   - design-basierte Simulation
   - Einfluss des Stichprobendesigns auf Methoden

Synthetische Populationen. Warum?

1) Viele interessante Variablen: nur durch Surveys
   - Anreicherung von Populationen mit Survey-Information
2) Simulationsstudien zur Evaluierung und Entwicklung von Methoden
   - design-basierte Simulation
   - Einfluss des Stichprobendesigns auf Methoden

Synthetische Populationen. Warum?

STATISTIK AUSTRIA
Die Informationsmanager

3) (Agent-based) Mikrosimulationen

- z.B. Gesundheitsplanung, Krankeitsausbreitung, Klimawandel, demographische oder wirtschaftliche Veränderungen – Vorausschätzungen auf Individualbasis.
- Ausgangspunkt ist eine Population
- Zukunftsträchtiges Forschungsgebiet
- “Geliebt” von Managern und Ökonomen
- Unbeliebt bei Statistikern (Inferenz?)

Synthetische Populationen. Warum?

STATISTIK AUSTRIA
Die Informationsmanager

4) Daten für Forschung und Lehre

- Lehre, speziell Stichprobentheorie
- Gute Trainingsdaten
- Public-use Files
- Bessere Strukturdatensätze
- Methodenentwicklung in der Forschung
- Disclosure risk $\rightarrow 0$ (Datenschutz ✓)

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Close-to-reality Daten

STATISTIK AUSTRIA
Die Informationsmanager

- Größe von Regionen und Strata erhalten
- Randverteilungen und Interaktionen zwischen Variablen sollten korrekt sein
- Hierarchische und Cluster-Strukturen
- Datenschutz muss respektiert werden
- Keine reine Replikation von Daten
- Einige Randvert. = Populationsgrößen

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

Inputdaten

Die Wahl der Methoden hängt von der **vorhandenen Information** und vom der gewünschten **Qualität** ab.

- Eine Datenquelle, mehrere Datenquellen
- Stichproben, Zensus, Tabellen
- Bekannte Populationsgrößen
- Für einfache Strukturfiles oder hochqualitative Daten

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

Methoden: 1) Synthetic reconstruction

STATISTIK AUSTRIA
Die Informationsmanager

. . . basiert auf

- bedingten Wahrscheinlichkeiten.
- Wird in Kombination mit Stichproben-Kalibrierungsmethoden verwendet (IPF, IPU)

kalibriertes Sample/Pop $\rightarrow$ synth. Population

Methoden: 2) Kombinatorische Optimierung

STATISTIK AUSTRIA
Die Informationsmanager

Kombinatorische Optimierung sind Techniken

- zur Kalibration von Populationen: $\theta = \theta^{(synth.Pop.)}$
- Anreicherung von detailierter geographischer Information
- Methoden: **Simulated Annealing**, Genetische Algorithmen, ...

Methoden: 3) Modell-basierte Methoden

Modell-basierte Methoden zur Simulation von close-to-reality Populationen

- mit Hilfe von Regressionsmethoden
- Mehrstufiger Prozess
- Sequentieller Prozess

Mehr später ...

Kalibration einer Stichprobe

- $S_i = 1$ wenn i in Stichprobe, sonst 0
- $Y = \sum_{i=1}^{N} y_i$ mit Pop der Größe N.
  Horwitz-Thomposon estimator $\hat{Y}_d = \sum_{i:S_i=1} d_i y_i$, mit
  Design-Gewicht $d_i = 1/\pi_i$.
- Kovariable $\boldsymbol{x}$ der Stichprobe mit bekanntem Total
  $X = \sum_{i=1}^{N} x_i$, und $\sum_{i:S_i=1} d_i x_i \neq X$. Finde neue
  Gewichte w_i mit $\hat{Y}_w = \sum_{i:S_i=1} w_i y_i$ wobei
  $\sum_{i:S_i=1} w_i x_i = X$ und $\sum_{i:S_i=1} w_i = N$.

Komplexer zB bei mehr NB und bei Haushaltsdaten

Kombinatorische Optimierung

STATISTIK AUSTRIA
Die Informationsmanager

- $\forall j = 1, ..., m \mid \theta_i = \theta_i^{(synth)}$
- Vergrößerte Population inklusive 0/1 Vektor
- Austausch von 0en und 1en bis alle Bedingungen erfüllt
- Rechenaufwendig vor allem bei Haushaltsinformation
- Target Swapping $\searrow$ Konvergenz (schneller)

Vergleich von Methoden

Vorteile	Nachteile
synthetic reconstruction	
▶ schnell ▶ einfach	▶ stetige Daten: no ▶ Haushaltsstr.: no
modellbasiert	
▶ Multivariate Struktur	▶ Rechenzeit: multinom ▶ Sequenz. $\rightarrow$ Resultat

Vergleich von Methoden

Vorteile	**Nachteile**
Kombinatorische Optimierung	
▶ Daten kalibriert	▶ Rechenzeit
Copulas	
▶ schnell	▶ komplexe Datenstrukturen: no

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

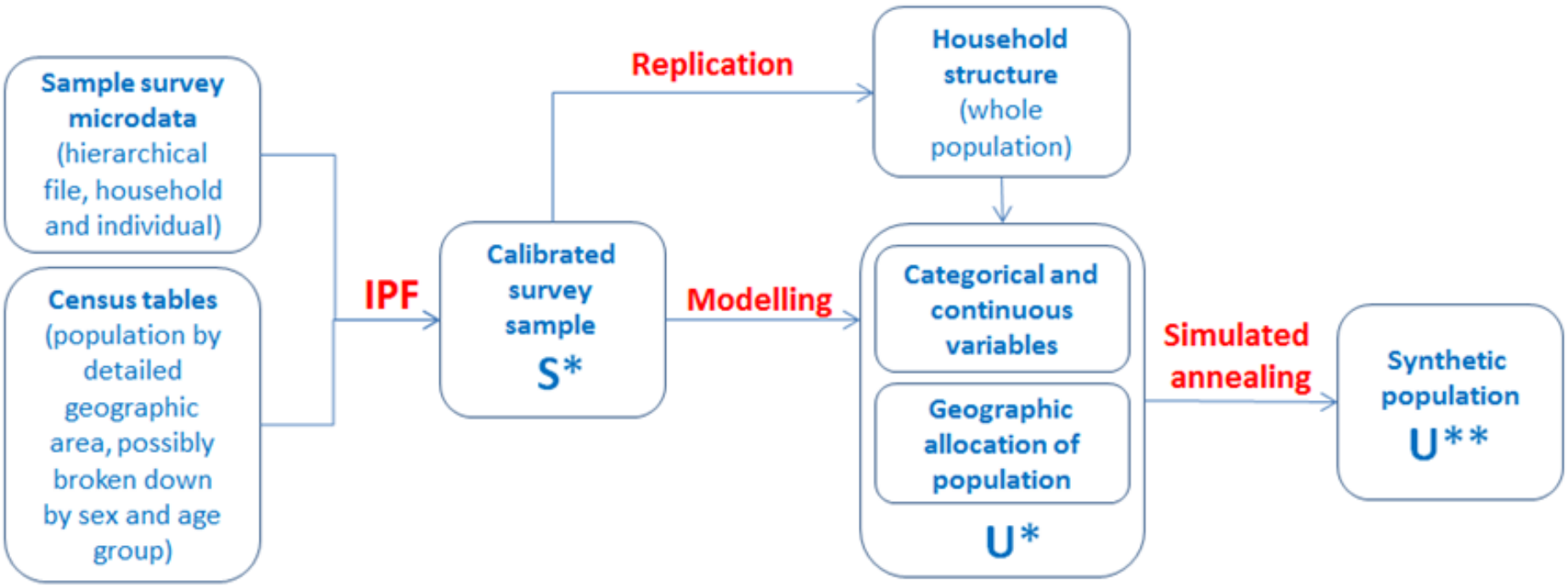
Methoden

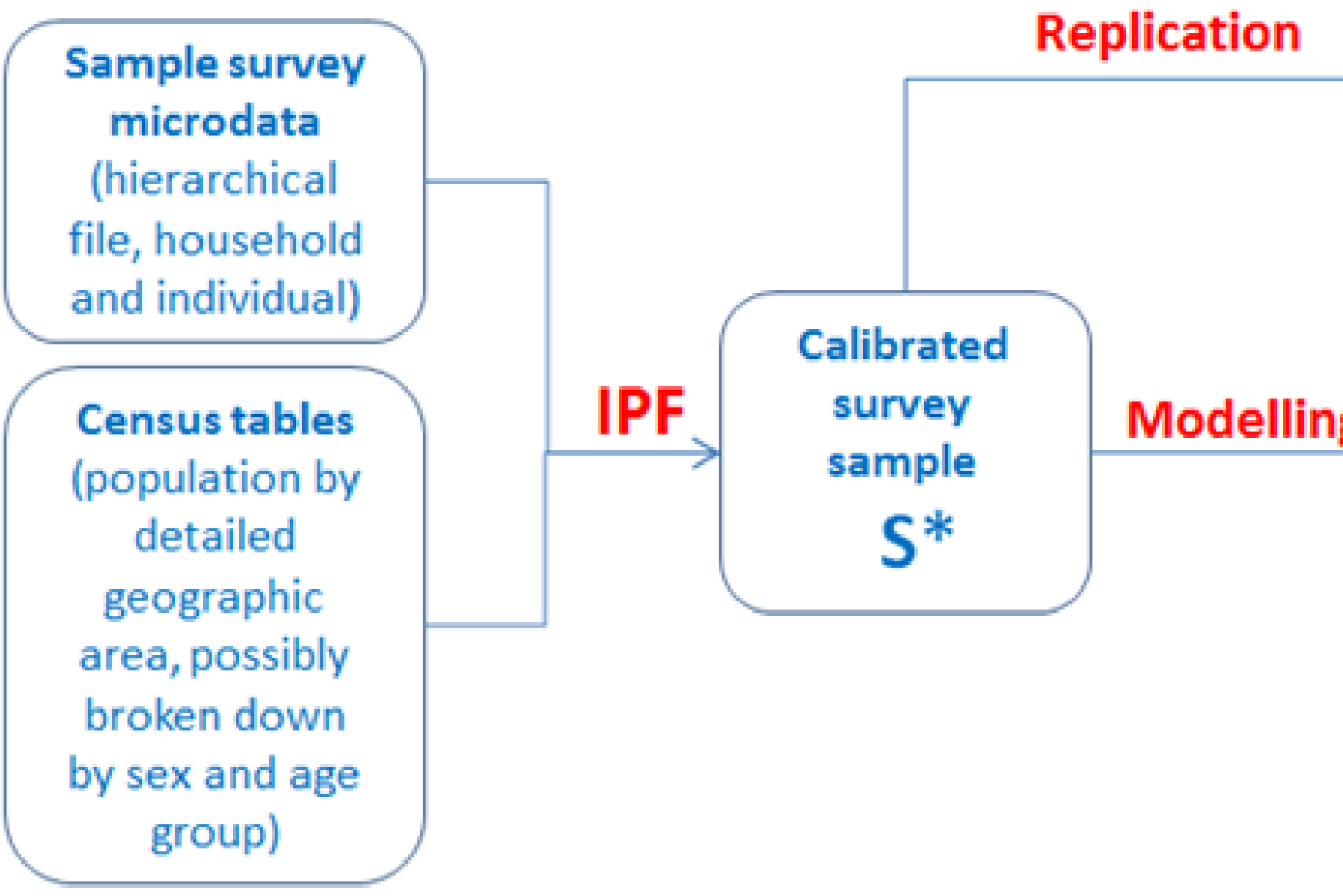
Modellbasierter Ansatz

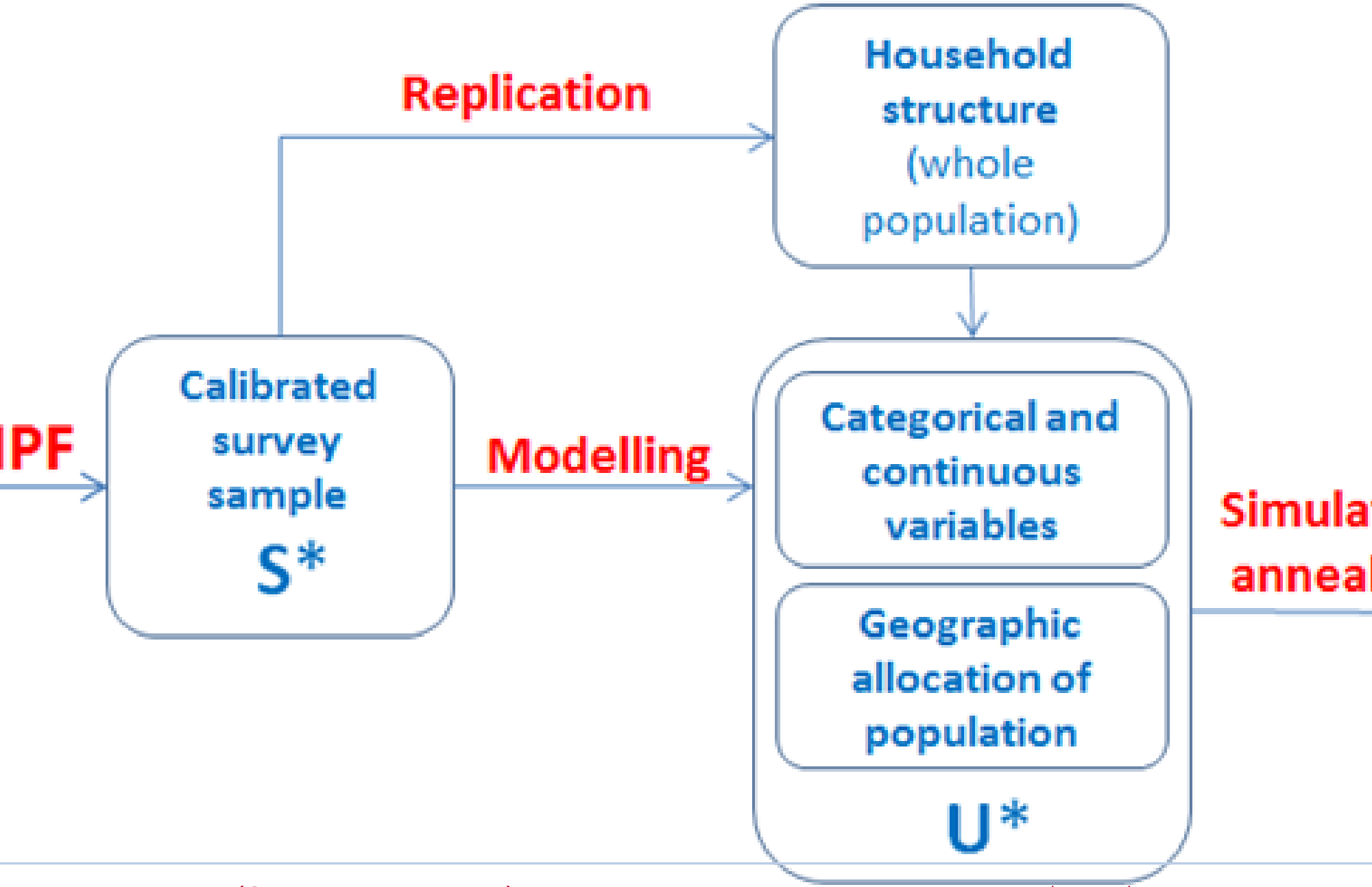
Package simPop

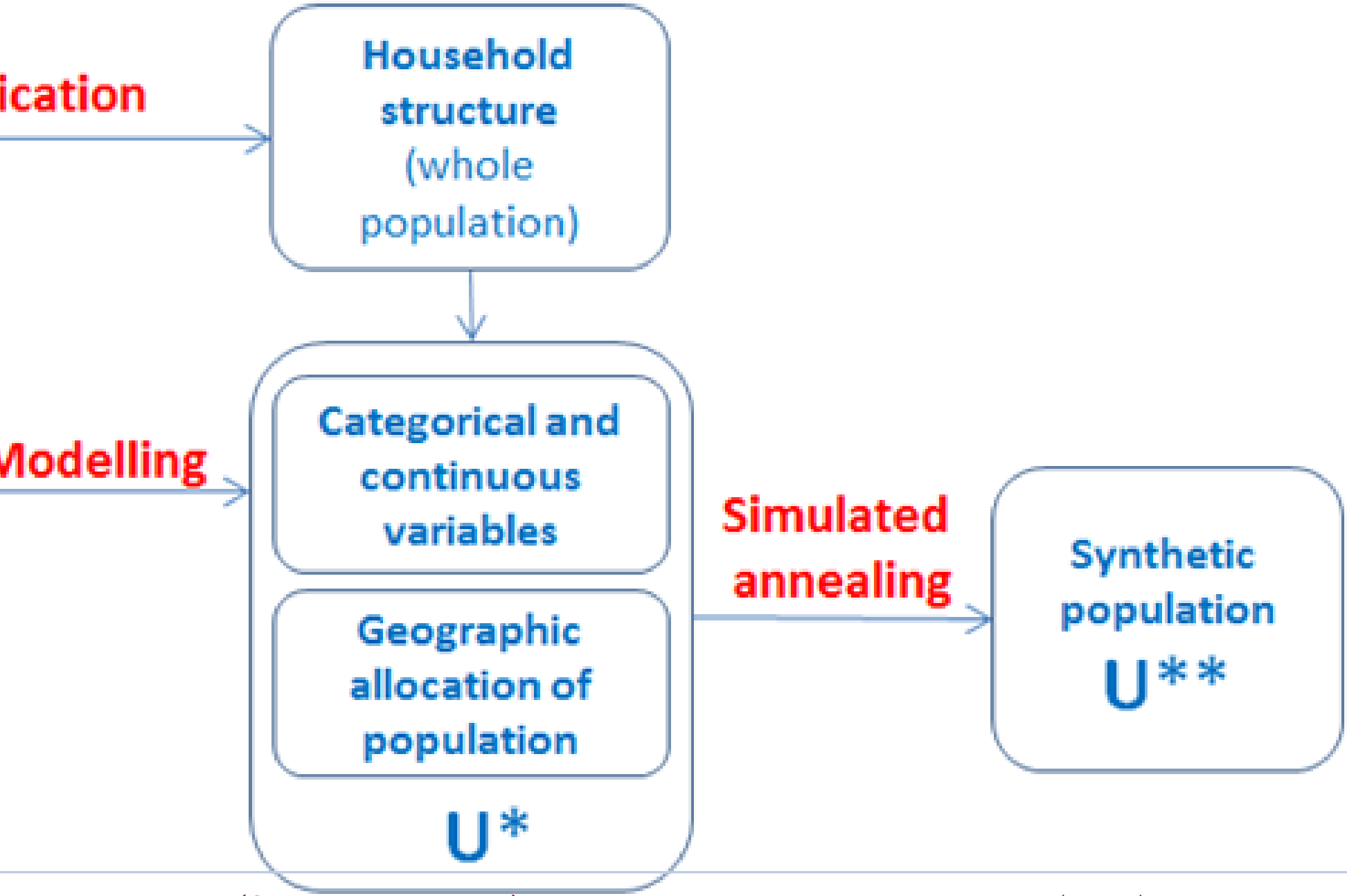
Beispiel: EU-SILC (vereinfacht)

Model-based approach: Vereinfachter Workflow




Replication
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*
Modellin


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


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

Model-based approach: Workflow

1. Erstellen der Haushaltsstruktur
2. Simulation von kategorischen Variablen
3. Simulation von stetigen Variablen
4. (Aufteilen von Komponenten)

- Stratifizierung um Heterogenitäten wiederzuspiegeln - Berücksichtigung der Stichprobengewichte

Modellbasiert: Strukturfile

Anwendung:

Haushaltsdaten mit Personeninformation

- **Haushaltstruktur** (core-variables): unabhängig in jeder Kombination von Haushaltsgröße und Strata
- # Haushalte: HT-Schätzung
- So wenig Variablen als nötig
- ZB Alter $\times$ Region $\times$ Geschlecht ($\forall$ Strata & HH)

Simulation kategorischer Variablen

- Input: Stichprobendaten
- Modell: Variable $\sim$ Kovariablen
- Regressionsparameter auf Pop übertragen

Neue Variable: Schätzung auf Stichprobe

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

$\longrightarrow$ Designmatrix zur Vorhersage von $\boldsymbol{x}_{j+1}$. Interaktionen berücksichtigen.

$\longrightarrow$ Schätzung der β's (Multinomiale Regression, Naivebayes,

2-Schritt-Verfahren, Regressionsbäume ...)

Neue Variable: Vorhersage auf Population

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

Bemerkung: Erwartungswerte werden nicht genommen $\longrightarrow$ aus Vorhersageverteilung ziehen

Simulation kategorischer Variables

- **Regressionsmethoden** angewandt auf Stichprobe
- Regressionskoeffizienten aus **Stichprobe** verwenden um Variable auf **Population** zu simulieren.
- Ziehen aus Wahrscheinlichkeiten der jeweiligen Kategorien auf individualer Basis.

Varianten: random Forests, ctrees, multinom, bedingte Wahrscheinlichkeiten

Simulation kategorischer Variables

- **Regressionsmethoden** angewandt auf Stichprobe
- Regressionskoeffizienten aus **Stichprobe** verwenden um Variable auf **Population** zu simulieren.
- Ziehen aus Wahrscheinlichkeiten der jeweiligen Kategorien auf individualer Basis.

Varianten: random Forests, ctrees, multinom, bedingte Wahrscheinlichkeiten

Simulation stetiger Variablen

- ähnliche Vorgangsweise entweder mit
  - multinomialen Model und Zufallsziehungen aus Kategorien
  - 2-Schritt Verfahren für semistetige Variablen

Zufallsfehler (Noise) durch Ziehen aus den Residuen hinzufügen.

Simulation stetiger Variablen

- ähnliche Vorgangsweise entweder mit
  - multinomialen Model und Zufallsziehungen aus Kategorien
  - 2-Schritt Verfahren für semistetige Variablen

Zufallsfehler (Noise) durch Ziehen aus den Residuen hinzufügen.

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

R package simPop

- Alle erwähnten Methoden & mehr
- Entwickelt von mehreren Insitutionen (Statitik Austria, TU Wien, ...)
- Objekt-orientiert (S4)
- effizient programmiert, auch für große Daten
- paralleles Rechnen automatisch

Warum benötigen wir synth. Populationen?

Was sind close-to-reality Daten?

Inputdaten

Methoden

Modellbasierter Ansatz

Package simPop

Beispiel: EU-SILC (vereinfacht)

EU-SILC

Mit Echtdaten siehe Templ, Spiess, Bergeat, and Meindl (2016b), Bergeat, Templ, and Spiess (2016)

```
library("simPop")
data("eusilcS")
origData <- eusilcS
origData$rb050 <- origData$rb050 * 100
## number of households (household ID: db030):
length(unique(origData$db030))


## [1] 4641
```

specifyInput()

```
inp <- specifyInput(origData,
                    hhid = "db030",
                    hhsize = "hsize",
                    strata = "db040",
                    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
##  ---------------
```

Kalibrierung Sample

```
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
```

Kalibrierung:

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

Haushaltsstruktur

```
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
```

Kategorische Variablen

```
synthP <- simCategorical(synthP,
                regModel = "available", # also for formulas
                additional = c("pl030", "pb220a"),
                method="multinom")
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
```

Stetige Variablen

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

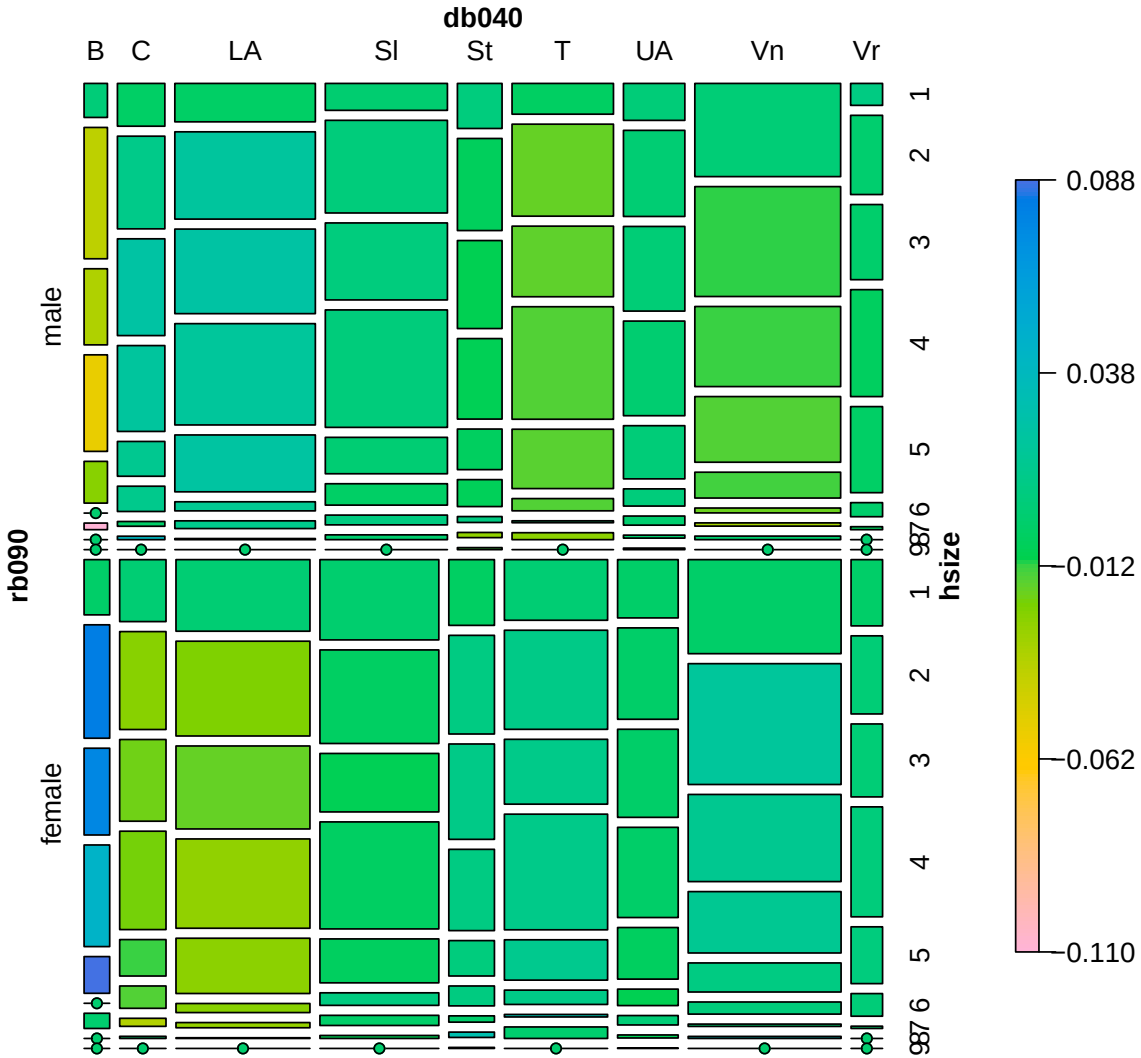
```
##
## --------------
## synthetic population  of size
##  8504755 x 11
##
## build from a sample of size
## 11725 x 19
## --------------
##
## variables in the population:
## db030,hsize,age,rb090,db040,pid,weight,pl030,pb220a,netIncomeCat,netIncome
```

Utiltiy

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

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

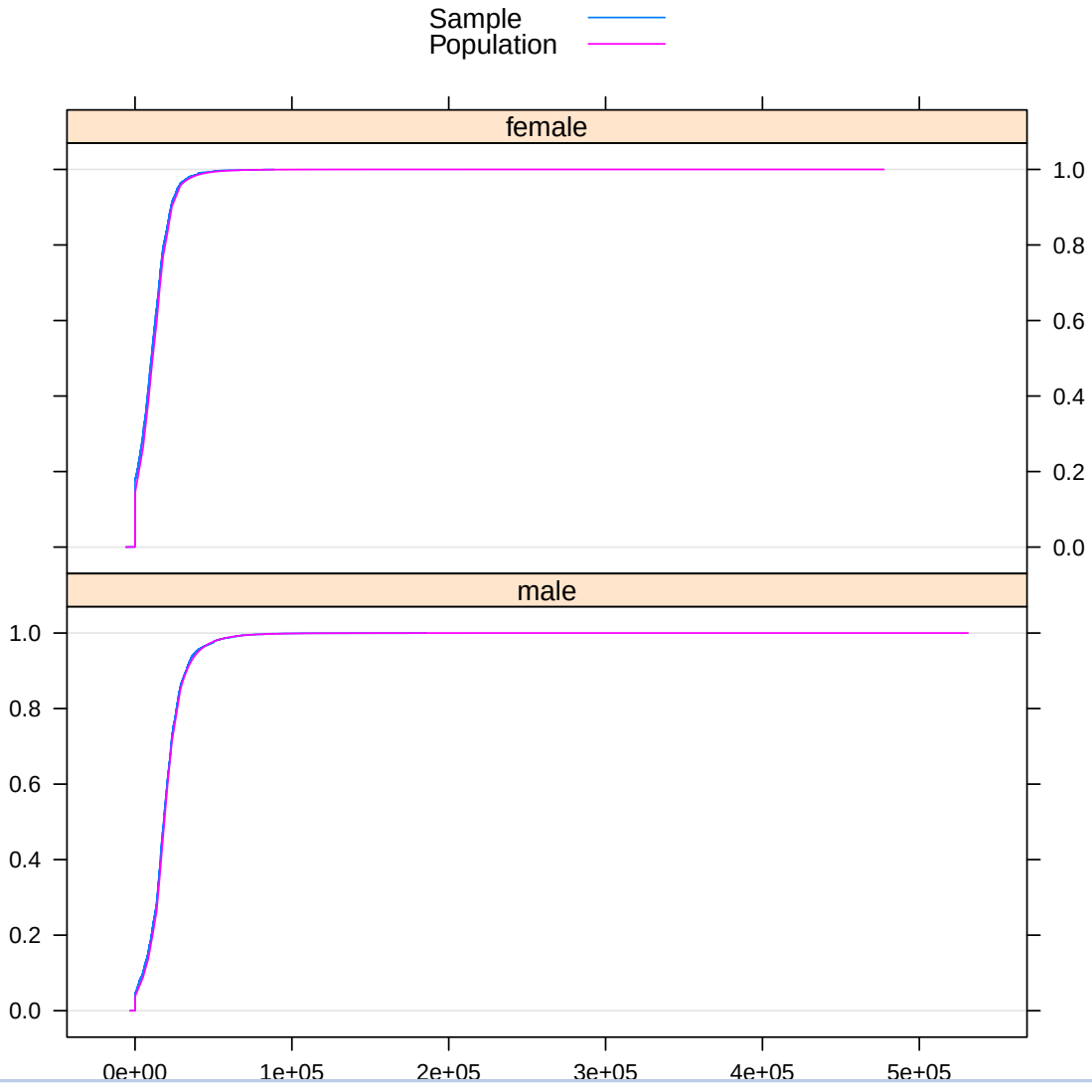
Utiltiy


db040
B
C
LA
Sl
St
T
UA
Vn
Vr
rb090
male
female
hsize
1
2
3
4
5
6
7
8
9
0.088
0.038
0.012
0.062
0.110

Utiltiy

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

Utiltiy


Sample
Population
female
male
1.0
0.8
0.6
0.4
0.2
0.0
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ◯ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ◯ Official Statistics (married!)

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ◯ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ◯ Official Statistics (married!)

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ◯ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ◯ Official Statistics (married!)

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ⊚ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ⊚ Official Statistics (married!)

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ◎ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ◎ Official Statistics (married!)

~~Vorurteile~~ Vorteile

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ⊚ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ⊚ Official Statistics (married!)

~~Vorurteile~~ Vorteile

STATISTIK AUSTRIA
Die Informationsmanager

- (Wir haben doch Populationen im Haus!)
  Anreicherung, Surveys, Weitergabe, Testen Methoden
- (Andere arbeiten auch nicht mit synth. Daten!)
  falsch & synth. Daten zukunftsträchtig
- ~~Synthetische Daten sind keine Echtdaten!~~
  Eigentlich doch. Sogar mehr (Populationen)
- ~~Mit synth. Daten verlieren wir Glaubwürdigkeit!~~
  Forscher ⊚ Official Statistics
- ~~Wir haben wichtigeres zu tun!~~ Reputation in Zusammenarbeit mit Forschern und open-access Data
- ~~Steckenpferd der Wissenschaft!~~ Notwendig für Wissenschaft

Forschung ⊚ Official Statistics (married!)

Referenzen

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. ISSN 1613-981X. doi: 10.1007/s10260-011-0163-2. URL http://dx.doi.org/10.1007/s10260-011-0163-2.

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

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.