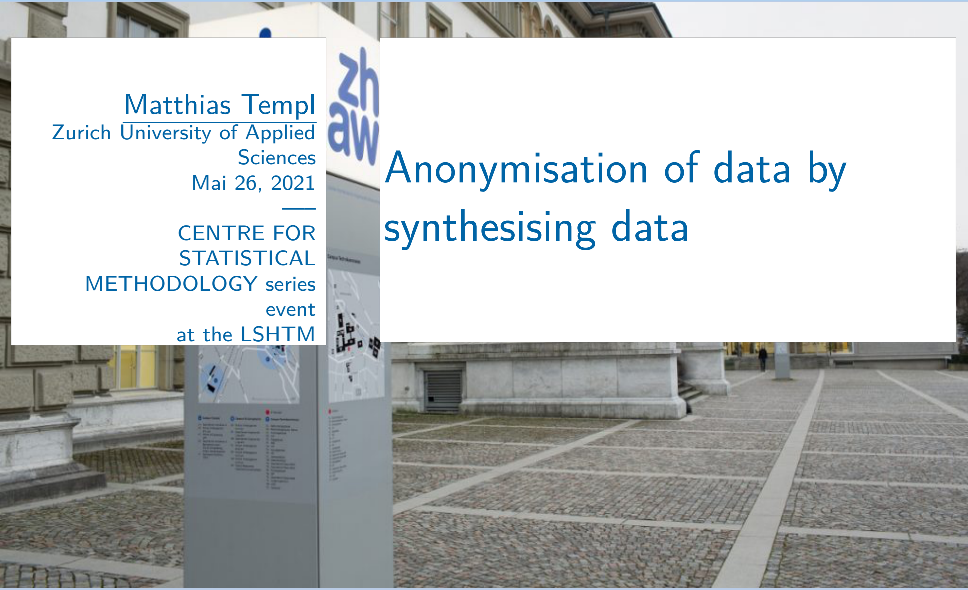


Matthias Templ
Zurich University of Applied Sciences

Mai 26, 2021

CENTRE FOR STATISTICAL METHODOLOGY series event at the LSHTM

# Anonymisation of data by synthesising data

# Motivation

Interview: Is synthetic data the key to healthcare clinical and business intelligence?

> *"synthetic data is now so popular [...]. Instead, almost any situation where real-world healthcare data is used can and probably is being represented with synthetic data. That allows for the low-cost, low-burden testing environment that then can be validated using real-world data." (Robert Lieberthal)*

# Overview

Introduction

What is meant by close-to-reality/realistic data?

Model-based simulation methods

Software simPop

Example

# Different concepts for anonymization

- **Traditional anonymization** to anonymize data under the paradigm of high data utility for scientific or public-use files
- **Remote execution**, **secure lab**, and **remote access**
- **Query servers** with perturbed aggregated output (differential privacy, secondary cell suppression, cellKey method, ...), e.g. for aggregated information in dashboards
- Black box methods to receive predictions on *test data* on sensitive variables without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated or centered learning** (PATE, ...)
- **Synthetic data**
  - as training data in machine learning
  - as *twin* data for the public or sharing within an organization
  - for augmented data or in form of population data

# Different concepts for anonymization

- **Traditional anonymization** to anonymize data under the paradigm of high data utility for scientific or public-use files
- **Remote execution**, **secure lab**, and **remote access**
- **Query servers** with perturbed aggregated output (differential privacy, secondary cell suppression, cellKey method, ...), e.g. for aggregated information in dashboards
- Black box methods to receive predictions on *test data* on sensitive variables without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated or centered learning** (PATE, ...)
- **Synthetic data**
  - as training data in machine learning
  - as *twin* data for the public or sharing within an organization
  - for augmented data or in form of population data

# Different concepts for anonymization

- **Traditional anonymization** to anonymize data under the paradigm of high data utility for scientific or public-use files
- **Remote execution**, **secure lab**, and **remote access**
- **Query servers** with perturbed aggregated output (differential privacy, secondary cell suppression, cellKey method, ...), e.g. for aggregated information in dashboards
- Black box methods to receive predictions on *test data* on sensitive variables without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated or centered learning** (PATE, ...)
- **Synthetic data**
  - as training data in machine learning
  - as *twin* data for the public or sharing within an organization
  - for augmented data or in form of population data

# Different concepts for anonymization

Zürcher Hochschule für Angewandte Wissenschaften

zh aw School of Engineering

- **Traditional anonymization** to anonymize data under the paradigm of high data utility for scientific or public-use files
- **Remote execution**, **secure lab**, and **remote access**
- **Query servers** with perturbed aggregated output (differential privacy, secondary cell suppression, cellKey method, ...), e.g. for aggregated information in dashboards
- Black box methods to receive predictions on *test data* on sensitive variables without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated or centered learning** (PATE, ...)
- **Synthetic data**
  - as training data in machine learning
  - as *twin* data for the public or sharing within an organization
  - for augmented data or in form of population data

# Different concepts for anonymization

- **Traditional anonymization** to anonymize data under the paradigm of high data utility for scientific or public-use files
- **Remote execution**, **secure lab**, and **remote access**
- **Query servers** with perturbed aggregated output (differential privacy, secondary cell suppression, cellKey method, ...), e.g. for aggregated information in dashboards
- Black box methods to receive predictions on *test data* on sensitive variables without access to *training data*
    - **Differential privacy** to noise output/predictions
    - **Federated or centered learning** (PATE, ...)
- **Synthetic data**
    - as training data in machine learning
    - as *twin* data for the public or sharing within an organization
    - for augmented data or in form of population data

# The traditional way of anonymising data

1) **RISK** Measurement of risk

- Sample or population? Micro data or tabular data?
- Other data sources with overlapping pop. to match with?
- Determination of a so-called **disclosure scenario**.
- Quantify the individual risk (of each person)

2) Anonymisation

- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?
- Perturbation of original data to lower the disclosure risk.

3) QUALITY Measurement of the utility

- Global procedures or data-specific comparisons?
- Which analysis results are of central interest?

# The traditional way of anonymising data

1) RISK Measurement of risk

- Sample or population? Micro data or tabular data?
- Other data sources with overlapping pop. to match with?
- Determination of a so-called **disclosure scenario**.
- Quantify the individual risk (of each person)

2) Anonymisation

- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?
- Perturbation of original data to lower the disclosure risk.

3) Measurement of the utility

- Global procedures or data-specific comparisons?
- Which analysis results are of central interest?

# The traditional way of anonymising data

1) RISK Measurement of risk

- Sample or population? Micro data or tabular data?
- Other data sources with overlapping pop. to match with?
- Determination of a so-called **disclosure scenario**.
- Quantify the individual risk (of each person)

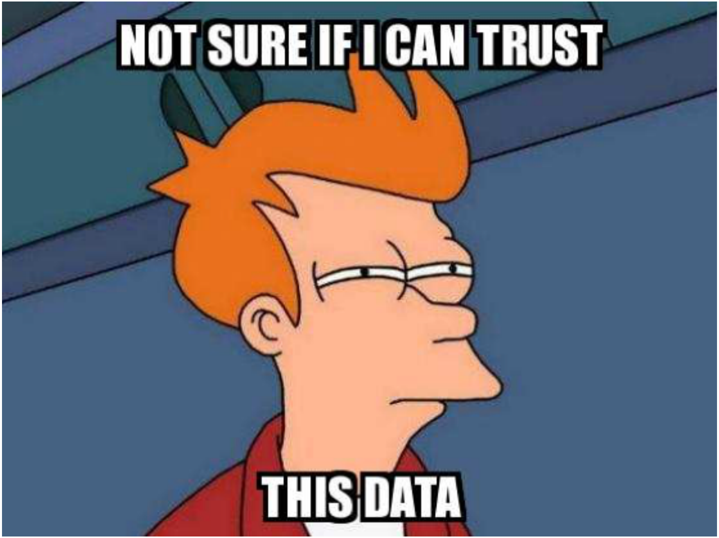
2) Anonymisation

- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?
- Perturbation of original data to lower the disclosure risk.

3) Measurement of the utility

- Global procedures or data-specific comparisons?
- Which analysis results are of central interest?

# Preconception



# Preconception

Unluckily, often some preconceptions are present related to synthetic data, some of them:

- We have a lot of data and do not need synthetic data/populations
- Others don't work with synthetic data
- Synthetic data are not real/true data
- Synthetic data → credibility loss
- We have more important issues to do
- Just a hobby from science in a dreaming spire

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

# Preconception

Unluckily, often some preconceptions are present related to synthetic data, some of them:

- We have a lot of data and do not need synthetic data/populations
- Others don't work with synthetic data
- Synthetic data are not real/true data
- Synthetic data → credibility loss
- We have more important issues to do
- Just a hobby from science in a dreaming spire

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

# Synthetic data: The holy grail?

- *“Synthetic data are as-good-as-real”* (Mostly AI)
- *“Preserves the statistical properties of your data”* (Statice)
- *“Maintains the stat. properties of real data.”* (replica analytics)
- *“The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original”* (Synthesized)
- *“a solution to generate synthetic twin datasets with the same statistical properties as the original data.”* (Diveplane)
- *“Why use real (sensitive) data when you can use synthetic data?” (Syntho)*

Ups, an outlier:

- *“Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data.”* (data-analysis OG)

# Synthetic data: The holy grail?

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

# Synthetic data: The holy grail?

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

# Synthetic data: The holy grail?

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

# Synthetic data: The holy grail?

Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

# Synthetic data: The holy grail?

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

# Synthetic data: The holy grail?

- *"Synthetic data are as-good-as-real"* (Mostly AI)
- *"Preserves the statistical properties of your data"* (Statice)
- *"Maintains the stat. properties of real data."* (replica analytics)
- *"The stat. properties of data are preserved [...] impossible to distinguish whether data is synth. or original"* (Synthesized)
- *"a solution to generate synthetic twin datasets with the same statistical properties as the original data."* (Diveplane)
- *"Why use real (sensitive) data when you can use synthetic data?" (Syntho)*

Ups, an outlier:

- *"Synthetic data are useful for a few applications, like data sets for training and education, simple struct. open-data with low disclosure risk, or augmenting real data."* (data-analysis OG)

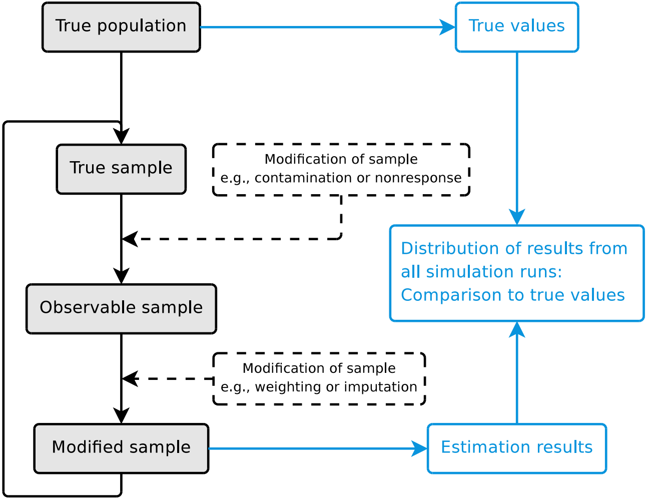
# Why synthetic data?

1) **Augment** data/populations with interesting variables from different sources
2) **Simulation studies** for the evaluation and development of methods
   - complex (design-based) simulation studies in survey methodology
   - influence of sampling designs on methods and results

# Why synthetic data?

1) **Augment** data/populations with interesting variables from different sources
2) **Simulation studies** for the evaluation and development of methods
    - complex (design-based) simulation studies in survey methodology
    - influence of sampling designs on methods and results

# ad 2) design-based **simulation studies**


True population
True values
True sample
Modification of sample
e.g., contamination or nonresponse
Observable sample
Distribution of results from all simulation runs: Comparison to true values
Modification of sample
e.g., weighting or imputation
Modified sample
Estimation results

# Why synthetic data?

3) For agent-based- and/or **micro-simulation**
   - e.g. health planning, spread of diseases, climate change forecasting, forecasting demographic and economic changes – all on individual (micro-level) basis
   - Starting point is a population of all individuals at time $T_0$
   - Hot topic in research. “Loved” by managers and ecometricians.

4) **Public-use data** for research, education, and the public

5) Because the **disclosure risk** → 0 (confidentiality issues ✓) (Templ and Alfons 2010)

6) As training data for remote exectution

(7) (As training data for machine learning methods)

Introduction

What is meant by close-to-reality/realistic data?

Model-based simulation methods

Software simPop

Example

# Close-to-reality data

- socio-demographic-economic structure of persons and strata need to be reflected
- marginal distributions and interactions between variables should be represented correctly
- hierarchical and cluster structures have to be preserved
- certain marginal distributions must exactly match known values
- data confidentiality must be ensured
  - no replication of units (e.g. using a bootstrap approach)

# First applications (historically)

- Clarke, Clarke, Birkin, Rees und Wilson (1984) simulated a population from aggregated data for the British (**health**) care organisation.
- Estimation of the **demand** of water (Clarke and Holm, 1987; Williamson, Birkin and Rees, 1998)
- From 1998 onwards a lot of applications such as
  - **health planning** (Brown and Harding, 2002; Tomintz, Clarke, and Rigby, 2008), (Smith, Pearce, and Harland, 2011),
  - **transportation** (Beckman, Baggerly, and McKay, 1996; Barthelemy and Toint, 2013)
  - **environmental planning** (Williamson, 2002).
- Evaluation and comparison of estimators and methods in DACSEIS, EUREDIT, AMELI, ..., research projects European level

Introduction

What is meant by close-to-reality/realistic data?

Model-based simulation methods

Software simPop

Example

# (Much) Too simplistic:

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

-
  ```
  X <- T %*% t(B) + E # component model
  ```
- simple model-based approaches
- to simulate data with the help of model-based imputation methods suggested by Rubin (1993) and many other authors
- use of simple replication methods (bootstrap, ...)
- using Copulas to simulate multivariate data
- simple use of deep learning methods

**For complex data, all these methods are too simplistic**

# Synthetic reconstruction methods

Original data → synth. ("twin") data

- Based on conditional probabilities estimated from original data
- Sequentially, categorical variables are simulated (`sample(...,` **prob**=`...)`)
- Non-observed categories or non-observed combinations of labels from several variables are not simulated (drawback)
- Only suitable for categorical data
- May be used in combination with sample calibration methods (IPF, IPU: iterative procedures for fitting marginal distributions)

# Model-based methods

... for the simulation of close-to-reality survey data or populations

- **multi-phase** process and sequential process
- special use of **regression** and **classification** (and in general machine learning up to deep learning) methods

Additionally needed

- **Calibration methods** to calibrate on known population characteristics
- **Combinatorial optimization methods** for the calibration of populations
- Tools to deal with special data problems, such as
  - **heaping** (e.g. age heaping)
  - **imputation methods** for missing values

# Model-based methods

... for the simulation of close-to-reality survey data or populations

- **multi-phase** process and sequential process
- special use of **regression** and **classification** (and in general machine learning up to deep learning) methods

Additionally needed

- **Calibration methods** to calibrate on known population characteristics
- **Combinatorial optimization methods** for the calibration of populations
- Tools to deal with special data problems, such as
  - **heaping** (e.g. age heaping)
  - **imputation methods** for missing values

Introduction

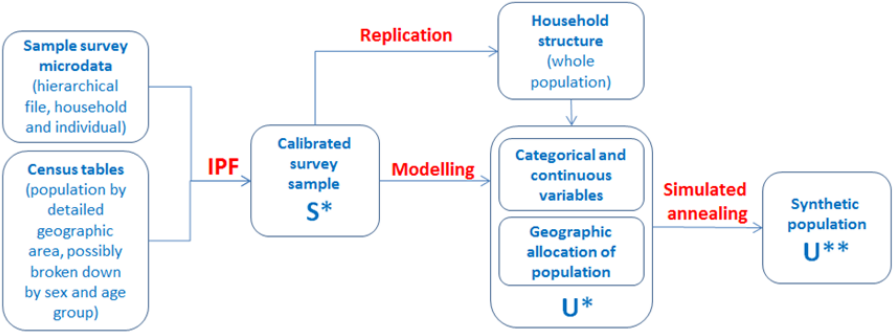
What is meant by close-to-reality/realistic data?

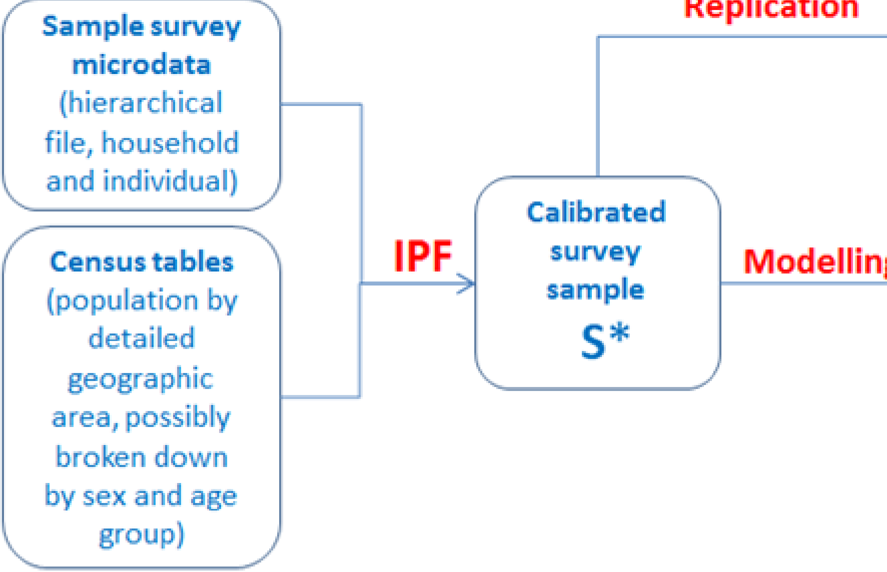
Model-based simulation methods

Software simPop

Example

# Model-based approach: simplified workflow


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



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 $S^*$

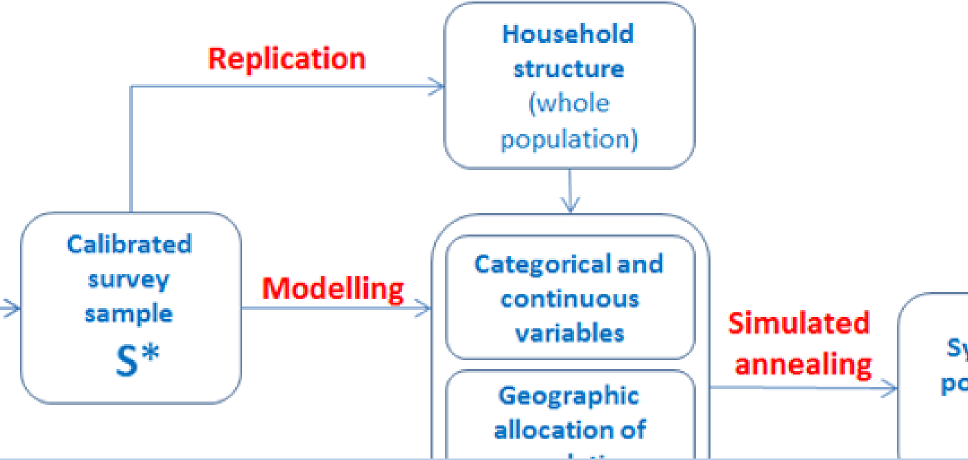
Modellin

# Calibration of a survey sample

- electronic health data may be subject to measurement error because of factors such as data entry errors and lack of documentation by physicians, non-responses and sampling bias.
- in case known population characteristics are known
- to achieve representativness (or at least reduce potential bias)
- Solution: iterative proportional fitting methods (**raking**)
- Cluster-Structures (e.g. persons in households), **iterative proportional updating**

# Socio-demographic base structure of patients/persons

Special structures such as patients in ambulances/hospitals, persons in households in health surveillance data, ...



Replication

Household structure (whole population)

Calibrated survey sample S*

Modelling

Categorical and continuous variables

Geographic allocation of

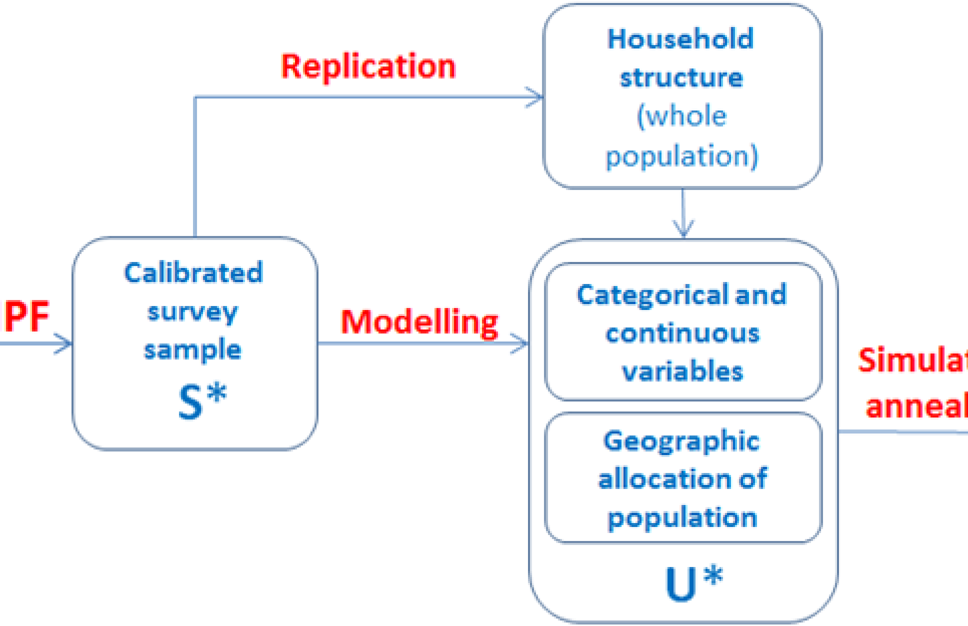
Simulated annealing

Sy
po

# Replication - model-based: Structure

Example: persons in households data

- **household structure** (core-variables): independently for every combination of household size and strata
- number of households: Horwitz-Thompson estimation
- for confidentiality issues, use only few variables for the structure
- e.g. age × region × gender (∀ strata & households)


Replication
Household structure (whole population)
PF
Calibrated survey sample S*
Modelling
Categorical and continuous variables
Geographic allocation of population
U*
Simulat
anneal

# Model-based approach: workflow

After setting up the household structure, additional variables are simulated using regression models:

1. simulation of categorical variables
2. simulation of (semi-) continuous variables
3. (simulation of compositions, e.g. parts on costs)

- stratification to reflect heterogeneity
- account for sampling weights (if any)
- account for missing values (if any)

# Estimation on the sample (sketch)

$$\textit{sample}\quad \boldsymbol{S} = \left(\begin{array}{ccccccc} \multicolumn{4}{c}{\overbrace{\hphantom{x_{1,1}\quad x_{1,2}\quad \cdots\quad x_{1,j}}}^{\text{“predictors”}}} & \overbrace{\hphantom{x_{1,j+1}}}^{\text{response}} & \multicolumn{2}{c}{\overbrace{\hphantom{x_{1,j+2}\quad \cdots}}^{\text{rest}}} \\ x_{1,1} & x_{1,2} & \cdots & x_{1,j} & x_{1,j+1} & x_{1,j+2} & \cdots \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & x_{2,j+1} & x_{2,j+2} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & x_{n,j+1} & x_{n,j+2} & \cdots \end{array}\right)$$

- design matrix to model $\boldsymbol{x}_{j+1}$
- Models of any complexity can be specified for each variable.
- estimation of the parameters, the “$\boldsymbol{\beta}$’s”, using linear models, robust models, multinomial regression, naive bayes, 2-step-approaches, regression trees, ctrees, xgboost, random forests, ann’s, ...

# Prediction of the population (sketch)

$$\textit{population} \quad \boldsymbol{U} = \begin{array}{c} \begin{array}{cc} \overbrace{\hspace{14em}}^{\hat{\boldsymbol{\beta}}\times \text{``pred.''} \approx} & \overbrace{\hspace{5em}}^{\hat{\mathbf{x}}_{j+1}} \end{array} \\ \begin{pmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \cdots & \hat{x}_{1,j} & \hat{x}_{1,j+1} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \cdots & \hat{x}_{2,j} & \hat{x}_{1,j+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{n,2} & \cdots & \hat{x}_{N,j} & \hat{x}_{1,j+1} \end{pmatrix} \end{array}$$

- We don't take the expected values, but draw from predictive distributions to account for model uncertainties.

# Model-based simulation of new variables

Categorical variables:

- methods: multinomial regression, naive-bayes, 2-step-approaches, classification trees, random forests, xgboost, artificial deep neural networks, ...

Continuous or semi-continuous variables

- multinomial model and draw from categories
- robust or ordinary least squares methods, glm's
- two-step approach for semi-continuous variables
- xgboost, random forests, artificial deep neural networks, ...

Random noise to account for model uncertainties is added to the fits/predictions by draws from the residuals or from the (normal) distribution of the residuals, or by dropout in ann's, etc.

# Model-based simulation of new variables

Categorical variables:

- methods: multinomial regression, naive-bayes, 2-step-approaches, classification trees, random forests, xgboost, artificial deep neural networks, ...

Continuous or semi-continuous variables

- multinomial model and draw from categories
- robust or ordinary least squares methods, glm's
- two-step approach for semi-continuous variables
- xgboost, random forests, artificial deep neural networks, ...

Random noise to account for model uncertainties is added to the fits/predictions by draws from the residuals or from the (normal) distribution of the residuals, or by dropout in ann's, etc.

# Model-based simulation of new variables

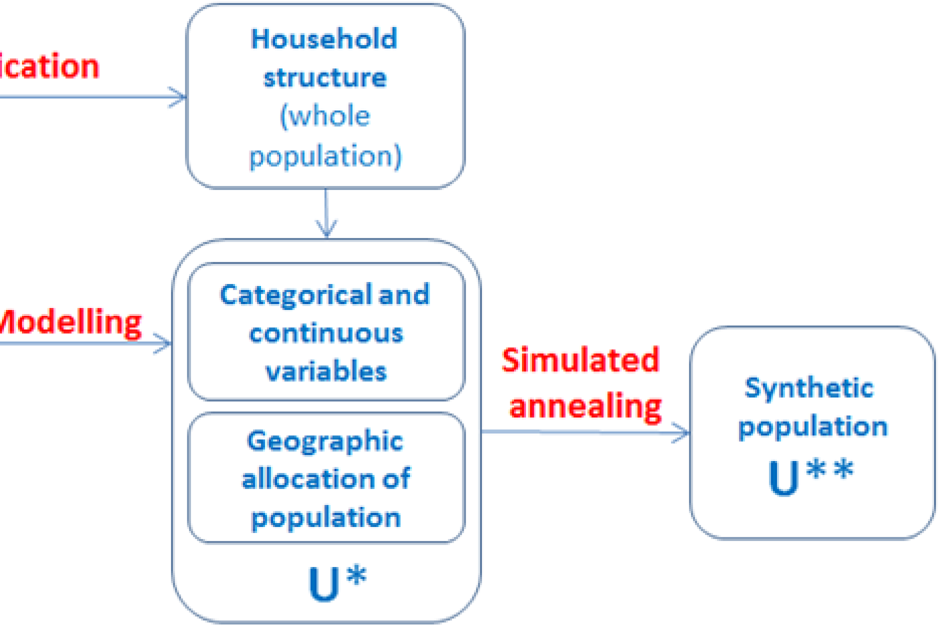
Categorical variables:

- methods: multinomial regression, naive-bayes, 2-step-approaches, classification trees, random forests, xgboost, artificial deep neural networks, ...

Continuous or semi-continuous variables

- multinomial model and draw from categories
- robust or ordinary least squares methods, glm's
- two-step approach for semi-continuous variables
- xgboost, random forests, artificial deep neural networks, ...

Random noise to account for model uncertainties is added to the fits/predictions by draws from the residuals or from the (normal) distribution of the residuals, or by dropout in ann's, etc.


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

# Combinatorial optimization

In case population data are simulated (and not only survey data). These techniques can be used to

- calibrate synthetic populations to receive consistent estimates for known marginal distributions (*swapping*, *target swapping*)
- *add finer geographical levels*
- Methods: **simulated annealing**, genetic algorithms, ...

Note: the price of simulating populations instead of sample data is (almost) zero, and one can always draw a sample from a population (if needed).

# Combinatorial optimization

In case population data are simulated (and not only survey data). These techniques can be used to

- calibrate synthetic populations to receive consistent estimates for known marginal distributions (*swapping*, *target swapping*)
- *add finer geographical levels*
- Methods: **simulated annealing**, genetic algorithms, ...

Note: the price of simulating populations instead of sample data is (almost) zero, and one can always draw a sample from a population (if needed).

# Untold stories

- adding finer geographical information
- methods against age heaping or heaping effects in continuous variables (e.g. rounded medication costs)
- imputation of item non-responses within the procedures
- evaluation of the disclosure risk and quality/utility of the synthetic population

Introduction

What is meant by close-to-reality/realistic data?

Model-based simulation methods

Software simPop

Example

# simPop history

- Theory: EU-FP7 project *AMELI* (Templ et.al, 2011)
- Software **simPopulation** (depricated)
- Software **simPop** (Templ, Meindl, Kowarik, and Dupriez 2017):
  - *Methods and tools for the generation of synthetic populations* (World Bank Project-No. 1129231)
  - *Synthetic populations and microsimulation* (World Bank Project-No 7177468)

Do not confuse with the synthpop R package, because of its similar name.

# R package simPop

- all mentioned methods & (much) more
- strictly object-oriented (S4 class implementation)
- efficiently programmed, can be used for huge data sets
- parallel computing is applied automatically
- enhanced “documentation” (published in the Journal of Statistical Software)
- last developments were supported by funds from the World bank

# Example

Using real-world data and simulating about 500 variables for several countries (Templ, Spiess, Bergeat, and Meindl 2016, Bergeat, Templ, and Spiess 2016) based on EU-SILC
New:

- Simulating health related variables such as chronic disease, general health condition, ... and covariates such as crime in neighborhood, leisure activities, schools nearby, economic status, occupational code, making do with net income, etc.
- Result is a synthetic data from the whole population simulated from survey data
- Can be used for micro-simulation (covid), education, training, open-data, ...

# Example

- We will only sketch how to use `simPop` (actual code is longer)
- We use public open-data to stay reproducible (with the price that the simulation of health related variables are not shown)

# Example

```r
library(simPop) # call simPop in R
# number of persons
nrow(origData)

## [1] 11725

# number of households (household ID: db030):
uniqueN(origData$db030)

## [1] 4641
```

# specifyInput()

```
inp <- specifyInput(origData,  # sample survey
             hhid = "db030",   # cluster ID (if any)
             strata = "db040", # by region
             weight = "rb050") # sampling w. (if any)
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
```

# Calibration of the survey sample

```r
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
```

Calibration:

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

# Calibration of the survey 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
```

## Calibration:

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

# Simulating the structure

```
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
```

# Categorical variables

```r
synthP <- simCategorical(synthP,
             regModel = "available", # or formulas
             additional = c("pl030", "pb220a"),
             method = "multinom") # ctree, randomForest, ...
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
```

# Continuous variables

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

# Health-related variables

## Sketch - not reproducible from slides (reason: privacy)

```r
synthP <- simCategorical(synthP,
  additional=c("health condition"),
  method = "multinom",
  regModel = formula("~ age + sex + crime + occupation +
                     parttime + makedoincome"))
synthP <- simCategorical(synthP,
  additional=c("chronic_disease"),
  method = "ann",
  regModel = formula("~ age + sex + health_condition + crime +
     occupation + parttime + makedoincome"))
synthP <- simCategorical(synthP, # better: simRelation
  additional=c("Blood group"),
  method = "xgboost",
  regModel = "basic")
```

# Calibrate population

- Again: census information to calibrate. External information (n-dimensional table) is available, e.g marginals on region gender economic status.
- We add these marginals to the object and calibrate afterwards

```
synthP <- addKnownMargins(synthP, totalsRG)
```

```
synthP <- calibPop(synthP)
```

```
 as also true for other functions, many parameters available,
here optional:  split=”db040”, temp=1, eps.factor=0.00005,
maxiter=200,  temp.cooldown=0.975,  factor.cooldown=0.85,
min.temp=0.001, verbose=FALSE
```

# Quality and disclosure risk of the synthetic data

Many utility measures possible, from **simple indicators**, to **visual comparisons**, to compare **point** and **variance estimates** for indicators, compare results from **models**.
The aim is always to **compare the sample information** or the information on known characteristics with results from the **synthetic data**.

- disclosure risk, see Templ and Alfons (2010)
- utility:
  - quality indicators: Templ (2017, 2015)
  - population based on EU-SILC: Alfons, Kraft, Templ, and Filzmoser (2011b), Bergeat et al. (2016), for employer-employee data: Templ and Filzmoser (2014)

We show two visual comparisons (can be done on finer detail)

# Quality/utility of the population

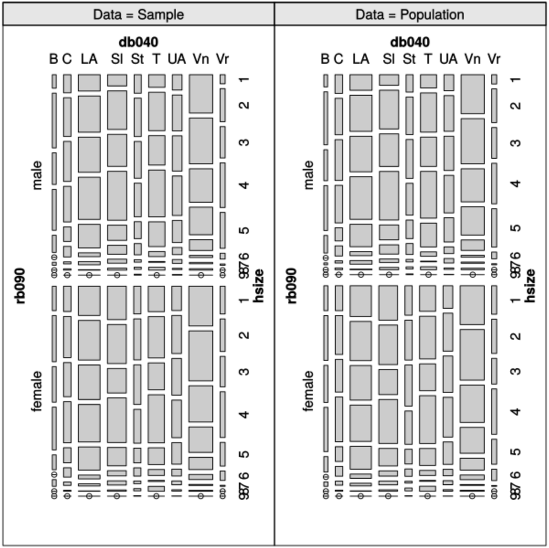
Tables (HT-(weighted) estimation):

```
tab <- spTable(synthP,
        select=c(”rb090”, ”db040”, ”hsize”))
```

Show frequencies visually:

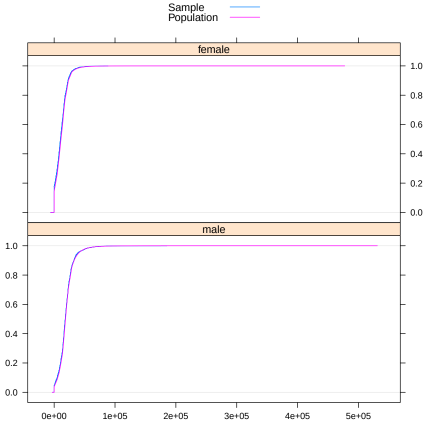
```
spMosaic(tab, # method = ”color”,
  labeling = labeling_border(
              abbreviate = c(db040 = TRUE)))
```

# Quality/utility of the population



# Quality/utility of the population

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



# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of simPop is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- simPop can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

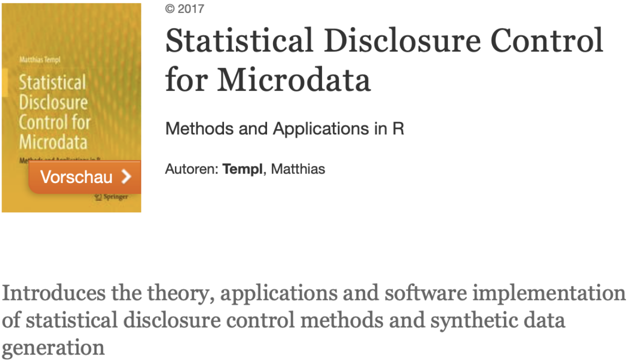
# Conclusions

Example was specific as each data set is specific.

- structure of original input is preserved
- margins of synthetic populations are calibrated
- all statistics can be almost precisely (but for many situations traditional anonymization is better)
- the synthetic data has very low disclosure risk
- code of `simPop` is quite efficient
- many other methods (classification trees, random forest, xgboost, ...) can be used
- population data as input for microsimulation
- open-access, public-use data, training data
- `simPop` can be used for many data sets in health sciences, different from the example shown (alternative software for less complex data structures available)

# Main references

1. http://www.springer.com/de/book/9783319502700


Matthias Templ
Statistical Disclosure Control for Microdata
Methods and Applications in R
Springer
Vorschau

© 2017

Statistical Disclosure Control for Microdata

Methods and Applications in R

Autoren: **Templ**, Matthias

Introduces the theory, applications and software implementation of statistical disclosure control methods and synthetic data generation

2. M. Templ, A. Kowarik, and B. Meindl. Simulation of synthetic complex data: The R-package simPop. Journal of Statistical Software, pages 1â39, 2016a.

# References

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, 2011a.

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, 2011b.

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

D.B. Rubin. Discussion: Statistical disclosure limitation. J Off Stat, 9(2):461–468, 1993.

M. Templ. Quality indicators for statistical disclosure methods: A case study on the structure of earnings survey. Journal of Official Statistics, 31(4):737–761, 2015.

M. Templ. Statistical Disclosure Control for Microdata. Methods and Applications in R. Springer, New York, 2017.

M. Templ and A. Alfons. Disclosure risk of synthetic population data with application in the case of EU-SILC. In J. Domingo-Ferrer and E. Magkos, editors, Privacy in Statistical Databases, volume 6344 of Lecture Notes in Computer Science, pages 174–186. Springer, Heidelberg, 2010.

M. Templ and P. Filzmoser. Simulation and quality of a synthetic close-to-reality employer–employee population. Journal of Applied Statistics, 41(5):1053–1072, 2014.

M. Templ, L. Spiess, M. Bergeat, and B. Meindl. Public use files for EU-SILC. SGA PUF Deliverable D3.1, Statistics Austria, 2016.

M. Templ, B. Meindl, A. Kowarik, and O. Dupriez. Simulation of synthetic complex data: The R package simPop. Journal of Statistical Software, 79(10):1–38, 2017. ISSN 1548-7660. doi: 10.18637/jss.v079.i10. URL https://www.jstatsoft.org/v079/i10.