Comparing fully synthetic Multiple Imputation with simPop

Klinck, Jannes
Sivakumaran, Thilipkumar
Slany, Emanuel

Otto-Friedrich-University Bamberg

further information:
www.github/emanuelsla/fully_synthetic_MI
branch analysis

07/03/2019

Overview

1. Outline
2. Synthetic data
   - Benefits
   - Drawbacks
3. Creating a fictive census
4. Fully synthetic Multiple Imputation
   - Basic idea
   - Pseudo algorithm
5. simPop
   - Package overview
   - Considered functions
   - Methods
6. Evaluation data utility
   - Graphical evaluation
   - Numerical evaluation
   - Household structure
7. Summary
8. Conclusion

Outline

- We will elaborate the basic benefits and drawbacks of using synthetic data.
- We will generate synthetic census data by using fully synthetic Multiple Imputation and the R package simPop from a fictive census.
- We will compare these approaches regarding their baseline idea and their data utility.
- We will examine the difficulty of using numeric measures for comparing synthetic datasets.

Synthetic data - Benefits

Definition: Synthetic data

The term synthetic data is the output of a procedure where we try to create a new fictive dataframe out of some given data.

- Synthesizing the data drives the disclosure risk towards zero. Therefore we focus on data utility in our research process.
- Synthetic data algorithms tend to meet the properties of the original data.
- Many synthetic data algorithms are ready-to-use.

Synthetic data - Drawbacks

- Synthetic data algorithms are mostly extremely computer intense.
- Some synthetic data algorithms generate implausible household structures.
- Results of some synthetic data algorithms are untraceable.

Creating a fictive census (1)

- We use the eusilcP dataset from the package simPop.
- We eliminate observations with not available values.
- We keep characteristic variables with different scales.
- We discard all observations with household id greater 1500.

Creating a fictive census (2)

The resulting data consists of 2854 observations and 8 variables and has the following structure:

Variable name	Scale of measurement
household_id	integer
region	nominal
household_size	integer
household_inc	numeric
age	integer
gender	factor
citizenship	factor
employment	factor

Table: Variables and their scales of our census

Fully synthetic Multiple Imputation - Basic idea

This approach goes back to Rubin (1993), who proposed to generate multiple, fully synthetic datasets for public release, so that no unit in the released data contains sensitive data from the actual unit in the population.

Our aim is to produce data that is similar to the actual data and has less restrictions in statistical inference.

The idea is related to Multiple Imputation, since the values, which we impute, are considered as missing.

This procedure leaves the data structurally unchanged but is very sensitive regarding the imputation model.

Fully synthetic Multiple Imputation - Pseudo algorithm (1)

Result: fully synthetic dataset

1. Start with the second variable of the dataset.
2. Fit a regression tree of this variable on the basis of the first variable.
3. In every leaf generate new values via Bayesian bootstrap.
4. Go to the next variable and fit a regression tree of this variable on the basis of the imputed variables. These are the columns on left-hand side.
5. Again, in every leaf generate new values via Bayesian bootstrap.
6. Repeat steps 4 and 5 till all variables are imputed. This is one resulting dataset.
7. Repeat the whole process n times to generate n independent datasets.

Fully synthetic Multiple Imputation - Pseudo algorithm (2)

The above approach goes back to Drechsler and Reiter (2010).
It originally covers partially synthetic modelling, but was extended to fully synthetic Multiple Imputation.

We added some extensions:

- a conditions vector to generate plausible values,
- a specification of the output options and
- specifications of split criterions.

simPop - Package overview

simPop: Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information

Tools and methods to simulate populations for surveys based on auxiliary data. The tools include model-based methods, calibration and combinatorial optimization algorithms.

Source: https://CRAN.R-project.org/package=simPop

simPop - Considered functions

Function name	Description
specifyInput	This function takes a data object as input and returns an s4-object. It allows to specify weights, household- and personal information and strata.
simStructure	This function simulates basic household variables by drawing samples from the data.
simCategorical	This function simulates the categorical variables.
simContinuous	This functions simulates the continuous variables.

Table: Executed functions from R package simPop with description

simPop - Methods (1)

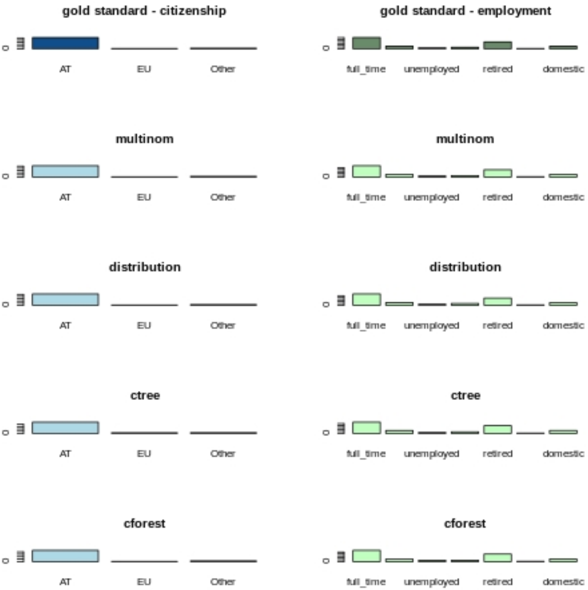
The functions

```
simCategorical()
simContinuous()
```

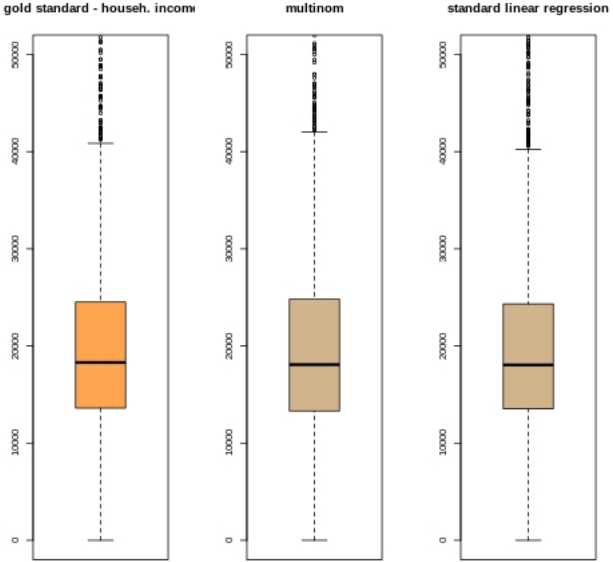
come with various methods.

We seek to choose the combination of methods, which synthesizes our census closest to the original data.
Therefore we synthesize our data with all available simPop-methods for these functions and decide on basis of a graphical evaluation.

simPop - Methods (2)



simPop - Methods (3)



simPop - Methods (4)

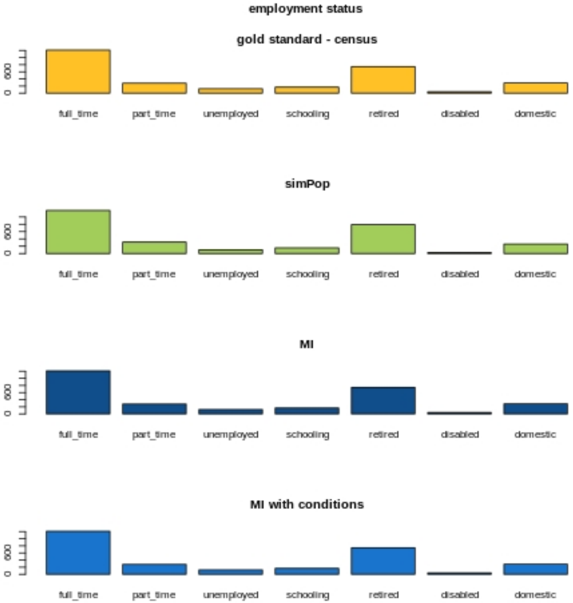
We decide to synthesize our data with the multinom-method in both functions.

Multinom method in simPop

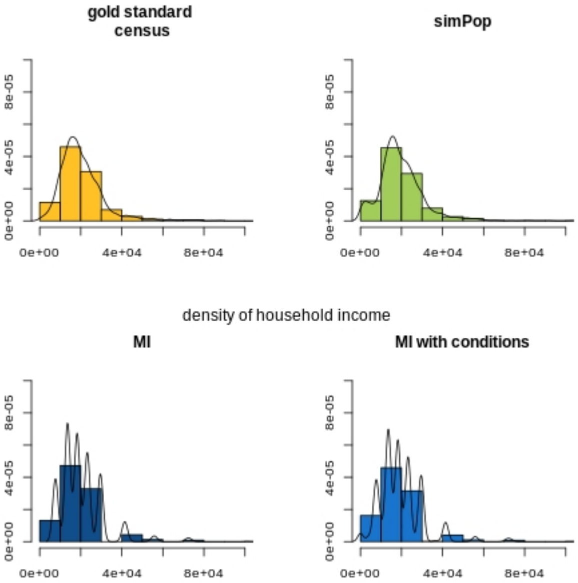
We create conditional probabilities using log-linear models and draw random draws from the resulting distributions.

Source: https://CRAN.R-project.org/package=simPop

Evaluation data utility - Graphical evaluation (1)



Evaluation data utility - Graphical evaluation (2)



Evaluation data utility - Numerical evaluation (1)

We decide to measure the overlap of the confidence intervals of selected parameters by the following formula:

$$I = \frac{U_i - L_i}{2(U_o - L_o)} + \frac{U_i - L_i}{2(U_s - L_s)}$$

Parameter	Description
U	Upper bound
L	Lower bound
U$_i$	Intersection of lower bounds
	(Same holds for L$_i$)
U$_o$	Observed lower bound
U$_s$	Synthetic lower bound

Table: Formula from Drechsler and Reiter (2009)

Evaluation of data utility - Numerical evaluation (2)

$$household\_inc = \beta_0 + \beta_1 age^2 + \beta_2 gender + \beta_3 region + \epsilon$$

Parameter	**simPop**	**synMI**	**synMi_cond**
age^2	0.92	0.96	0.66
gender	0.91	0.99	0.95
region	0.56	0.98	0.97

Table: Intersect overlap measure according Drechsler and Reiter (2009)

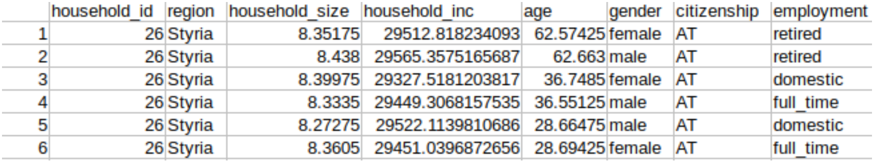
Evaluation of data utility - Numerical evaluation (3)

Further we review, if the relation between the variables is still intact after synthesizing the dataset, which is done by analyzing the correlations.

Dataset	Correlation
Gold	0.03
simPop	0.02
synMI	0.04
synMI_cond	0.06

Table: Correlations of household_inc and age

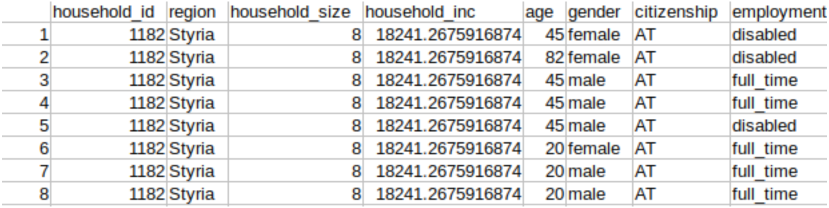
Evaluation data utility - Household structure (1)



	household_id	region	household_size	household_inc	age	gender	citizenship	employment
1	26	Styria	8.35175	29512.818234093	62.57425	female	AT	retired
2	26	Styria	8.438	29565.3575165687	62.663	male	AT	retired
3	26	Styria	8.39975	29327.5181203817	36.7485	female	AT	domestic
4	26	Styria	8.3335	29449.3068157535	36.55125	male	AT	full_time
5	26	Styria	8.27275	29522.1139810686	28.66475	male	AT	domestic
6	26	Styria	8.3605	29451.0396872656	28.69425	female	AT	full_time

Example: Household from synthetic data with Multiple Imputation

Evaluation data utility - Household structure (2)



	household_id	region	household_size	household_inc	age	gender	citizenship	employment
1	1182	Styria	8	18241.2675916874	45	female	AT	disabled
2	1182	Styria	8	18241.2675916874	82	female	AT	disabled
3	1182	Styria	8	18241.2675916874	45	male	AT	full_time
4	1182	Styria	8	18241.2675916874	45	male	AT	full_time
5	1182	Styria	8	18241.2675916874	45	male	AT	disabled
6	1182	Styria	8	18241.2675916874	20	female	AT	full_time
7	1182	Styria	8	18241.2675916874	20	male	AT	full_time
8	1182	Styria	8	18241.2675916874	20	male	AT	full_time

Example: Household from synthetic data with Multiple Imputation with conditions

Summary

- All methods perform well for synthesizing categorical variables.
- simPop creates better results for continuous variables.
- Multiple Imputation holds overall properties for continuous variables, but destroys density.
- Conditions in Multiple Imputation create a more plausible household structure, but contradict the general approach.
- Quality of estimates depend strongly on underlying method.

Conclusion

Overall simPop generates more appropriate synthetic dataframes.
But estimates hold their quality after synthesizing not generally.

Multiple Imputation meets general properties, but destroys densities and logical structures.
Often the attempt to prevent these problems generate even worse results.
This problem can maybe be fronted with pruning.

The remaining problem is to find numerical indicators, which compare whole synthetic datasets with the original dataset.

References

Drechsler, Reiter (2009)
Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata.
Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 589603.

Drechsler, Reiter (2010)
Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata.
Journal of the American Statistical Association, Vol. 105, No. 492, p. 1347-1357.

Meindl et al. (2010)
Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information.
https://github.com/statistikat/simPop

Rubin, Meng (1993)
Maximum likelihood estimation via the ECM algorithm: A general framework.
Biometrika, Vol. 80, pages 267-278.

Discussion

What are appropriate numerical measurements
for comparing data utility of different synthetic datasets?