Comparing fully synthetic Multiple Imputation with simPop

Jannes Klinck, Thilipkumar Sivakumaran and Emanuel Slany

Otto-Friedrich-University Bamberg

all supplementary material on:
www.github.com/emanuelsla/fully_synthetic_MI
branch: analysis

Abstract. The importance of disclosure control grows every month. The European Data Protection Regulation has catalyzed this development. Institutions and companies are looking for an appropriate way to store and publish data in a convenient, legal and fruitful manner.

The term "synthetic data" is referred to an output of procedures, which try to anonymize the data in an automatic way. These methods have some benefits. Mainly they are able to drive the disclosure risk towards zero. Downsides of these procedures are critical aspects like exponential computational effort.

We evaluate fully synthetic Multiple Imputation and the R package simPop regarding the properties and plausibility of the results for synthesizing a census. In order to do so, we create fictive census data based on the eusilcP dataset.

Fully synthetic Multiple Imputation goes back to missing data analysis, since we treat the values to impute as missing. The R package simPop comes with many different methods. We use samples with conditional probabilities from log-linear models to synthesize the data. The basic result is that beyond univariate distributions of categorical data the usage of simPop is beneficial.

To further underline this result, we use the survey dataset eusilcS. The synthetic data of simPop results in tremendously more precise estimates in comparison to the Multiple Imputation approach and the original survey.

A side note is given to the problem, which occurred as we searched for numerical measurement procedures to compare the synthetic data with the original dataset. These methods are often position specific and therefore not suitable for synthetic data analysis.

Keywords: Disclosure Control · R · Multiple Imputation · simPop.

1 Introduction

Especially in the recent scientific development the research on disclosure control reached high importance. Since the procedure of anonymizing data by simply creating a new dataset with the same properties as the original data appears to be one of the most practical solutions for disclosure control, it is therefor necessary to thoroughly compare the different methods for creating said synthetic data.

Synthetic data in general drag the disclosure risk towards zero. This is achieved by creating new observations, which do not try to directly mimic observations from the original data. Due to this fact, the comparison between newly created and original data becomes increasingly more complex, which in turn complicates the evaluation of the disclosure risk. Since this issue can not be sufficiently resolved in the scope of this paper, we abstain from any further discussion on this topic.

While in theory we simply create algorithms, which compile new datasets with the same properties as the original dataset, we soon encounter drawbacks in practice. Many of these algorithms are enormously computer intensive for larger number of observations. As we will see some algorithms generate implausible household structures. And especially for end-users these methods can be considered as a black box, since the results are not generally reproducible. We will see how these attributes behave for the fully synthetic Multiple Imputation algortihm and the simPop approach while they synthesize a fictive census, which will be created in the next section. We will state the general principle of both methods and evaluate the results regarding mathematical properties and plausibility. Additionally we are going to validate the results of both methods for survey data as input.

Lastly we will give a short conclusion on the comparison.

2 Create a fictive census

To maintain a gold standard for the comparison of both approaches, we create fictive census data. For that we use the eusilcP dataset from the R package simPop. We eliminate all observations, which have at least one missing value. To decrease computation time, we omit all observations with a household identification number over 1500. We apply additional computation to ensure a logical structure within the census data. The resulting census has 2854 observations and 8 variables. Since we want to keep basic information on households and have different scales of measurement for our variables, the census has the structure as displayed in **Table 1**. In the following we will use the variable names in the text.

3 Fully synthetic Multiple Imputation

This approach will create multiple, fully synthetic datasets, so that no unit in the synthesized data contains sensitive information from the census data. This approach goes back to Rubin (1993) and is, as the name already says, related to Multiple Imputation in missing data analysis. We will not go into detail at this point, since the derivation of Multiple Imputation methods is sometimes

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 1. Variables of the census and their scales

cumbersome. Just to connect the concept of Multiple Imputation with disclosure control: we treat the values, which we impute, as missing data. And since we do this for all data points, we result in fully synthetic data.
However the original algorithm, derived by Drechsler and Reiter (2010), covered partially synthetic Multiple Imputation. Their algorithm was adopted in former research and extended towards fully synthetic Multiple Imputation. The basic structure of the algorithm can be seen in **Algorithm 1**.

```
Data: Dataset to protect
Result: Fully synthetic dataset
for all desired repetitions do
    start with second variable
    while not all variables are synthetic do
        if numeric variable then
            run a regression tree
            insert buckets
            for all buckets do
                run a bootstrap
                save results
            end
        end
        else
            run a classification tree
            insert buckets
            for all buckets do
                run a bootstrap
                save results
            end
        end
    end
end
```

Algorithm 1: Fully synthetic Multiple Imputation

Simplified we start with the second variable in the dataset and run a regression tree on the first variable. Then we draw a Bayesian bootstrap in every leaf. After

that we go to the third variable and fit a regression tree on all variables on the left side. These are the first two variables. Again we apply a Bayesian bootstrap in every leaf. In this fashion we iterate over the whole dataset.
Bayesian bootstrap means that we draw multiple items based on the posterior. In this case the regression tree is considered as prior and the actual dataset as likelihood. We distinguish between a numerical case, where we use a regression model and a bootstrap with respect to the mean and a categorical case, where we use a classification model and a bootstrap based on the frequency table.
We made some additional extensions like output options, options for pruning and a conditions vector to generate more plausible results.

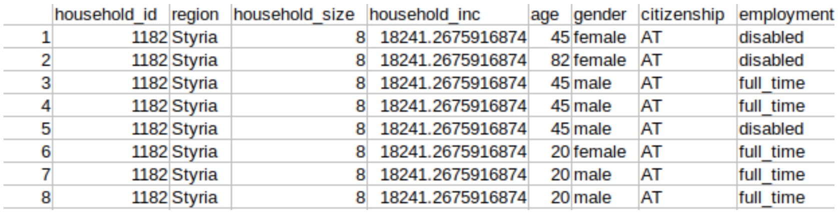
4 simPop

Whereas the Multiple Imputation approach stems from recent research, simPop is a substantial package for synthesizing data. It is available on CRAN and GitHub and comes with multicore processing, which is a major benefit in this resource-consuming field. It includes multiple functions and methods.
In our case we created an s4-object, which was originally meant to store multiple data objects. We simulated the basic household structure using random draws from the observed conditional distributions within the strata. At this stage basic household variables (age and gender) were passed to the function.
After creating the basic structure, we began simulating the categorical variables (employment and citizenship). Lastly we used the afore-mentioned simulated variables as basis for the estimation of the household income, where again multiple methods were provided for this simulation by the package. We investigated all opportunities and decide to use random draws from the distributions of log-linear models, since it seems to have minor benefits in comparison to the other methods.

5 Evaluation

In this section we will compare the results of synthesizing the data with fully synthetic Multiple Imputation with and without conditions with the simPop approach in relation to the fictive census.
As **Fig. 1** illustrates via investigating one household from the respective dataset, the conditions vector increases the plausibility in comparison to the standard Multiple Imputation procedure. However there still are unneglectable lacks in plausibility for example in the age structure.

Although all methods perform nearly equally good for synthesizing categorical variables, we see big differences in the quality of numeric variables between the procedures. This is stated in **Fig. 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

Fig. 1. Example household from fully synthetic Multiple Imputation dataset with conditions

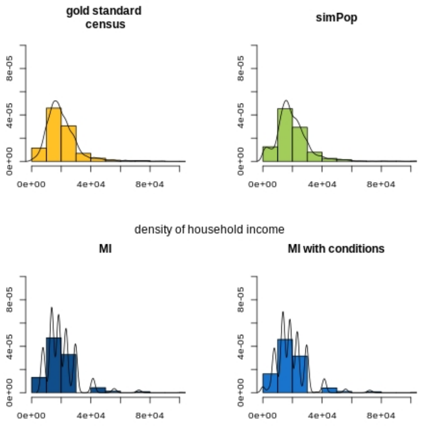


Fig. 2. Comparison of household_inc of methods in relation to the census

For further evaluation we use a basic linear regression model with the following structure:

$$household\_inc = \beta_0 + \beta_1 age^2 + \beta_2 gender + \beta_3 region + \epsilon \tag{1}$$

We estimate the confidence interval of the coefficients and measure the overlap with the confidence intervals from the census data. This is accomplished by the following method by Drechsler and Reiter (2009). The parameters are explained in **Table 2**.

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

Table 3 visualizes the result. It reinforces the above-mentioned argument that Multiple Imputation has major drawbacks for synthesizing numeric data in

Parameter	Description
U	Upper bound
L	Lower bound
U_i	Intersection of upper bounds (same holds for L_i)
U_o	Observed lower bound
U_s	Synthetic lower bound

Table 2. Notes on formula from Drechsler and Reiter (2009)

comparison to simPop. That simPop performs worse for categorical data is due to the fact that, because of pruning in the regression tree, Multiple Imputation only changes values with high disclosure risk. This is an effective solution.

Parameter	simPop	MI	MI with conds.
age^2	0.92	0.96	0.66
gender	0.91	0.99	0.95
region	0.56	0.98	0.97

Table 3. Intersect overlap measure according Drechsler and Reiter (2009)

6 Application to survey data

Until now we chose the usecase such, that we have a fair comparison of both methods. But since the simPop package can also synthesize survey data, we will change towards this field of use, although we know that it is not the intended usage for the fully synthetic Multiple Imputation algorithm.
As comparison criterion we reutilize the equation (2), which is explained in **Table 2**. We also reuse the linear regression model (1), but apply a log transformation on the income variable. As the underlying survey for this comparison stems from eusilcS, data on personal income becomes available and is chosen as the dependent variable in the regression over household income. We only compare the results of simPop and the Multiple Imputation method without conditions with the original survey data.

Table 4 shows the result. Numbers outside of the range between zero and one indicate no intersection between the confidence intervals. Fully synthetic Multiple Imputation clearly suffers a lot by including weights. The extend of which is being shown by the negative values produced by the intersect overlap measure, which indicates no overlap at all with the intervals of the regression parameters from the original survey data. Even if simPop also nearly misses the intersection area once, it is extremely beneficial especially in this usecase.

Parameter	simPop	MI
age^2	0.65	-3.43
gender	0.07	-4.3
region (median)	0.69	0.68

Table 4. Intersect overlap measure according Drechsler and Reiter (2009) on eusilcS

7 Conclusion

Without summing up all results again, we draw the following conclusion: The more complicated the underlying structure of the data, which we want to synthesize, is, the more beneficial it is to use simPop instead of fully synthetic Multiple Imputation. This holds for all points. Multiple Imputation performs worse in these settings in terms of general properties, plausibility and computation time. Further steps on this topic could be to generalize the conditions vector in the fully synthetic Multiple Imputation algorithm in a way that increases plausibility and meets the principle of Multiple Imputation. We should derive a numeric way for estimating the disclosure risk. Clearly synthetic data push the risk towards zero, but not all observations will have a zero disclosure probability. Lastly we seek for a way to derive one numeric criterion that compares the synthetic data with the original data. Known metrics mostly compare the synthetic and original data in a row by row manner, but since fully synthetic algorithms pay no regard to positioning in the dataset, these metrics are ill-suited in the given case. This calls for the creation of a new criterion that also suites synthetic data.

8 References

1. Drechsler, Reiter (2009):
   Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey.
   Journal of Official Statistics, Vol. 25, No. 4, 2009, pp. 589603.
2. 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.
3. Meindl et al. (2010):
   Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information.
   https://github.com/statistikat/simPop.
4. Rubin, Meng (1993):
   Maximum likelihood estimation via the ECM algorithm: A general framework.
   Biometrika, Vol. 80, pages 267-278.