Outline paper - Multiple imputation for statistical disclosure control: Creating synthetic datasets with *mice*

Anony*mice*d shareable data: Using *mice* to create multiply imputed synthetic datasets

Thom Volker, Gerko Vink & Stef van Buuren

20-08-2020

Table of Contents

# Introduction

Open science, including open data, has been marked as the future of science (Gewin, 2016), and the advantages of publicly available research data are numerous (Foster & Deardorff, 2017; Walport & Brest, 2011). Additionally, the fact that public funds are used for data collection results in increased demands for this data. However, research data is very often collected under the reassurance of privacy and confidentiality of the data, and simply anonimizing the data is not enough to fulfil these requirements (Ohm, 2009). Over the years, several techniques have been used to increase the confidentiality of the data, such as categorizing continuous variables, top coding values above a certain threshold or adding random noise (Drechsler, 2011). However, these methods may distort relationships between variables, reducing the data quality drastically.

An alternative has been offered by Rubin (1993), who, building on the framework of multiple imputation, proposed to release multiple synthetic datasets to the public. Conform this approach, all units that are in the population but not in the sample are treated as missing data. These values are imputed by means of conventional multiple imputation approaches, and simple random samples are drawn from the population, and released to the public. In practice, it is not required to impute the complete population, since random samples can be drawn from the sampling frame, so that only these sampled values have to be imputed. The imputed samples generated under this approach are labelled as *fully synthetic datasets*.

A second procedure to creating synthetic datasets has been proposed by Little (1993), named *partially synthetic datasets*, which yields imputing only those values that are at a high risk of being disclosive. These values can be complete variables that could be compared with publicly available datasets or registers, such as addresses, or it can be certain values that bear a high risk of being disclosive, say, income values above a certain threshold. Similarly to the fully synthetic data approach, multiple datasets containing synthetic values are released to the public, although in this instance, a subset of the data remains constant over the imputations, due to it’s non-identifying nature.

– – NOT SURE YET: CONTRAST FULLY AND PARTIALLY SYNTHETIC DATA – –

– – SINCE PARTIALLY SYNTHETIC DATA POOLING RULES DO NOT REQUIRE ADJUSTED VARIANCE ESTIMATES – –

Imputing synthetic values is conceptually similar to the imputations of missing values. That is, conform fully conditional specification, values are drawn iteratively from the posterior predictive distribution of the variable of interest, conditional on all other variables, for each variable separately, where the previously imputed values are taken into account in the imputation model for the variables that are imputed later on. ***ADD MORE INFO FCS***

– – CAN WE REGARD THE OBSERVED VALUES AS THE STARTING VALUES FOR THE IMPUTATIONS ? – –

The R-package mice (Van Buuren & Groothuis-Oudshoorn, 2011) implements the fully conditional specification approach for missing data. Since the imputation approaches for missing data and synthetic data do not differ in terms of the imputation process, mice can be used for creating synthetic data as well, which will be shown in more detail in the remainder of this paper. In the next section, the pooling rules that are required to obtain inferences from the synthetic datasets are discussed. Thereafter, it will be shown that creating confidence valid [TO DO: FULLY/PARTIALLY] synthetic data can be generated by means of mice. Also, it is shown that iterating over the conditional distributions in order to establish convergence is not required, due to the fact that new values are drawn directly from posterior distribution given the observed data.

# FULLY/PARTIALLY synthetic data

DISCUSS APPROACH IN MORE DEPTH

# Mice

* Discuss mice, predominantly the imputation algorithm in greater depth
* Procedure: observed values can be regarded as starting values, so that all variables can be used for the synthesis of all other variables.

# Simulations

## Methods

* Boys data - bootstrapped to obtain a population.
* explain simple synthetic data procedure
* explain imputation settings
* CART has been proposed by Reiter

## Results

## References

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation* (Vol. 201). Springer Science & Business Media.

Foster, E. D., & Deardorff, A. (2017). Open science framework (osf). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

Gewin, V. (2016). Data sharing: An open mind on open data. *Nature*, *529*(7584), 117–119. <https://doi.org/10.1038/nj7584-117a>

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*(2), 407.

Ohm, P. (2009). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, *57*(6), 1701–1778. Retrieved from <https://heinonline.org/HOL/P?h=hein.journals/uclalr57&i=1713>

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from <https://www.jstatsoft.org/v45/i03/>

Walport, M., & Brest, P. (2011). Sharing research data to improve public health. *The Lancet*, *377*(9765), 537–539.