PhD Research Proposal

About the Evaluation of Missing Data Methodologies

Author: Rianne Schouten

Supervisors: prof.dr. Stef van Buuren, dr. Gerko Vink
University of Utrecht, Department of Methodology and Statistics

February 27, 2018

1 Introduction

The occurence of missing data may be problematic in every situation where data is stored, processsed or analyzed. Missing data occurs in scientific research where datasets can be incomplete due to non-response or drop-out. And in an application-orientated field like data science available datasets suffer from incompleteness as well. In order to make sensible policy decisions based on reliable analysis results, a proper handling of the unobserved data is essential. Therefore, my research will focus on the evaluation of methods dealing with these missing data.

Of all possible methods for dealing with missing data, the most advantageous missing data method for a specific missing data problem is determined by - among others - the scientist's aim with the data. It is important to realize this aim could differ per discipline or per field. For instance, data scientists generally predict the value (or category) of an output variable. In this *prediction* domain, the goal of data analysis is to obtain the model with the most accurate predictions (James, Witten, Hastie, & Tibshirani, 2014). In contrast, scientific researchers aim at drawing valid statistical inferences about a population. More specifically, they intent to investigate how an output variable changes as a function of certain input variables. Here, the aim of data analysis is to explore the relation between variables. We call this statistical *inference* making (James et al., 2014).

It depends on the research goal of a given situation whether a missing data method is appropriate to deal with the missing values. For data scientists, fast and easy methods - such as deleting incomplete rows from the dataset (i.e. drop) or imputing the missing values with the mean or median of a feature - are appealing approaches. In addition, a data scientist may choose to use crossvalidation techniques to compare the performance of more than one missing data method. Eventually, the appropriateness of a missing data methodology is determined by its capability to accurately predict the output variable (Hastie, Tibshirani, & Friedman, 2009). In the world of inference making, on the other hand, missing data methods are evaluated by their ability to enable the finding of valid statistical inferences. Suitable missing data methods are methods that result in unbiased population estimates without disturbing the covariance structure of the data. In general, multiple imputation is considered to be a useful technique (Van Buuren, 2012).

Great part of missing data theory is written in the context of statistical inference making (Rubin, 1976, 1987; Schafer & Graham, 2002). As a consequence, we know a lot about subtle differences between missing

data methodologies, what to do in special cases such as surveys with small sample sizes and how to deal with missing data in particular types of analyses such as longitudinal data analysis and multilevel analysis. In addition, several authors have provided us with systematic comparisons of missing data methods (e.g. Peeters, Zondervan-Zwijnenburg, Vink, & Van de Schoot, 2015) and their software implementations (e.g. Horton & Kleinman, 2007). However, the translation of all this knowledge to situations where prediction is the main purpose of the analysis seems to occur only sporadically. For example, data science books use only 1 or 2 pages to describe missing data methods (Hastie et al., 2009; James et al., 2014). Remarkably, the few missing data methods that are discussed, are particularly those methods that scientific researchers consider inappropriate for most analyses (i.e. listwise deletion, mean imputation). Other places where data scientists may search for the knowledge they need, such as blogposts or data science journals, also rarely cover the topic of missing data.

With my research, I intent to form a bridge between two fundamentally different worlds that both have to deal with missing data. Data scientists could benefit from the large collection of scientific literature, provided that a sufficient translation and application of this knowledge is available. On the other hand, missing data methodologists may use the far-reaching experience of data scientists with machine learning techniques to further develop and implement missing data methods. In addition, data scientists have a very practical and standardized way of working. A standardized approach of missing data problems may be valuable for scientific researchers as well and the data science practice could serve as an example for the development of clear and logical algorithms.

2 Evaluation

As the title of this document points out, my research focuses on the evaluation of missing data methodologies. The few data science documents discussing the use of missing data methods evaluate the performance of these methods with so-called accuracy measures. Examples of these measures are the confusion matrix, precision and recall measures and the area under the ROC curve for classification tasks, and mean squared error (MSE) and explained variance (R^2) for regression tasks. It is easy to obtain accuracy measures with the `.metrics` module from Scikit-learn in Python (Pedregosa et al., 2011). Since the goal of data science is to predict the value or class of an output variable - thus, to obtain the best accuracy - researchers evaluate the performance of missing data methods with accuracy measures. For instance, Acuña and Rodriguez (2004, p. 6) used "the 10-fold crossvalidation estimates of the misclassification error." Apart from factors such as efficiency and computation time, the missing data method that gives the best *prediction accuracy* is considered to be the best way to deal with the missing values.

Another way to evaluate the performance of missing data methodologies is to calculate the *imputation accuracy*. Imputation accuracy refers to the extent to which imputation methods are able to find an imputation value close to the original, missing value. Again, accuracy measures such as MSE and misclassification rates are used. However, instead of comparing predictions with the true outcome values as happens in calculating the prediction accuracy, imputation accuracy is calculated by comparing the imputed values with the original values - the values that became missing. Logically, imputation accuracy can only be calculated when the complete, original dataset is available. This is the case in a simulation setting. For instance, Tang and Ishwaran (2016) use a combined measure of standardized RMSE and misclassification error to compare random forest algorithms. When a missing data method returns good measures of imputation accuracy, it means that the method reproduces the original data well. Therefore, comparing missing data methods by

their imputation accuracy may be a good way to evaluate their performance.

Because most missing data simulations are performed in the domain of statistical inferences, missing data methods are mostly evaluated with measures of statistical validity. Here, the interest is in knowing the extent to which a missing data method is able to give unbiased and efficient statistical estimates. For instance, with which missing data method is my linear regression model a good representation of the population? Evaluation measures that are often used are bias of a statistical estimate, coverage rate of the 95% confidence interval and the confidence interval width of that interval (Van Buuren, 2012).

3 Aims

The aim of my research project is to provide applied scientists with knowledge about missing data techniques. Properly solving missing data problems requires an understanding of the effect of missing data problems on the outcome of analyses. Because most missing data literature has focused on the domain of inference making, it is yet unclear to what extent the scientific knowledge can directly be utilized in the prediction domain. In addition, the large datasets and demanding computational techniques give that applied scientists often choose for fast and easy missing data methods such as listwise deletion or mean imputation. The ultimate goal of my project will be **to investigate to what extent applied scientists may rely on simple, easy and fast missing data methods and under which circumstances more complicated solutions for missing data are required?**

In order to reach this research aim, an underlying, more theoretical question has to be addressed: **While evaluating missing data methodologies, what do we consider to be the truth and why?** More specifically, should we use prediction accuracy, imputation accuracy, statistical validity or any other measure to evaluate the performance of missing data methodologies? According to Garcia-Laencina, Sancho-Gomez, and Figuiras-Vidal (2010, p. 280), "in classification tasks with missing values, the main objective of an imputation method must be to help to enhance the classification accuracy." However, Shah, Bartlett, Carpenter, Nicholas, and Hemingway (2014, p. 772) state that "better predictions do not mean better coverage of confidence intervals; it is important that imputation methods incorporate the correct amount of variation in order to produce unbiased estimates with correct coverage of confidence intervals" and Van Buuren (2012, p. 46) is also clear: "We cannot evaluate imputation methods by their ability to re-create the true data, or by their ability to optimize classification accuracy. Imputation is not prediction." With my research I intent to give numerical insights into the differences between evaluation methods. Overall, I aim at making the following concept key in my work: "A missing value treatment cannot be properly evaluated apart from the modeling, estimation, or testing procedure in which it is embedded" (Schafer & Graham, 2002, p. 149).

4 Chapters

4.1 Amputation

The evaluation of missing data methods is often done with simulation studies. Such simulation studies generally have four steps:

1. A complete dataset is acquired from practice or simulated from multivariate distributions.

2. The complete dataset is made incomplete.

3. The missing values are estimated and imputed by means of missing data methodologies.

4. Statistical measures are obtained for the original, complete dataset and after dealing with the missing values. A comparison of these measures gives an indication of the performance of the missing data method.

The advantage of using real, complete datasets is the complicated but realistical mixture of categorical and continuous variables. However, the specific characteristics of real datasets can make it hard to generalize findings to other datasets. Therefore, simulating complete datasets is considered to be a valuable and useful technique as well. A procedure where one complete dataset is sampled and considered to be the population of interest (Brand, Van Buuren, Groothuis-Oudshoorn, & Gelsema, 2003) is a useful approach because it gives that the effects of the sampling mechanism do not have to be taken into account when calculating the evaluation measures. As such, the focus of the simulation study can be on solving the missing data problem.

Part of a proper evaluation of missing data methodologies is the generation of realistic and legitimate missing data. We refer to the process where missingness is induced in complete data as *amputation*. I have started my project with the development and implementation of a multivariate amputation procedure (`ampute`: Schouten, Lugtig, Brand, & Vink, 2017). The availibility of `ampute` provides a way to systematically explore the effects of missing data problems on statistical analyses. Especially `ampute`'s ease to generate different missing data problems is valuable for testing realistic missing data situations. In the first chapter of my project, I will introduce the multivariate amputation procedure and provide evidence for prefering `ampute` over the traditional amputation methods.

4.2 Robustness

The second chapter of my research will focus on the robustness of missing data methods for missing data mechanisms. The difference between (completely) random and non-random missing data problems - MCAR, MAR and MNAR respectively - is a widely discussed topic in missing data theory (Rubin, 1976, 1987). When evaluating missing data methods, it is important to distinguish between these three types of missing data problems. In general, a missing data method such as multiple imputation will yield valid statistical estimates under MCAR and MAR missingness, and has the potential to deal with MNAR data (Rubin, 1987; Van Buuren, 2012). A simple and easy method such as listwise deletion can give unbiased estimates under MCAR but could cause trouble for statistical inferences under MAR and MNAR. In this chapter, I will interpret the outcomes of evaluation procedures in terms of missing data mechanisms. More specific, I will investigate whether there are circumstances where statistical inferences are similar under different missingness mechanisms, and whether the use of measures of prediction accuracy influences the interpretation of missingness mechanisms in the prediction domain. Let me explain these questions further.

First, I will investigate what the correlation between variables should be to justify the MAR assumption under MNAR. It is generally accepted that MNAR data are very common in practice. For example, scientist Mariana Simons-Nikolova explained to me that as part of the development of medical technology, Philips Research Europe invites patients to repeatedly visit a test clinic to measure certain health qualities such as blood pressure and heart rate. With these data, Philips is able to estimate and improve the performance of their technology. However, patients with acute or increased health issues may temporarily be prevented from visiting these test clinics and as a result, Philips' observed data contains the medical results from patients who are, on average, healthier than the total patient population. Now, if extra information about the differences between two patients groups is used, the quality of the analysis could be improved. In missing data terms,

you could say extra information enhances the covariance structure of the total dataset, and as such makes the MAR assumption more reliable. Collins, Schafer, and Kam (2001) showed that the inclusion of a variable that correlates either with the incomplete variable or with the missingness improves parameter estimation in the situation of MNAR. My research will build on this by showing the behaviour of estimates of the mean, variance and correlation of a bivariate normal distribution for different data correlations (Schouten & Vink, 2017).

Second, I will investigate the differences between MCAR, MAR and MNAR missingness for data science models where crossvalidation procedures are used to predict a certain outcome variable. To determine whether a missing data problem can be categorized as MCAR, MAR or MNAR, we have to think about the question: "What part of the dataset is incomplete?" On the one hand, answering this question requires determining whether the missing values are in the features and/or in the outcome variable. A data scientist aims to predict an output variable and therefore, it is quite common to drop records from the dataset with a missing value on the output variable. In particular, the incomplete record is considered to have *no label* (Hastie et al., 2009; James et al., 2014). However, what happens if the unlabeled records are distinct from the labeled records? What to do if you want to predict unlabeled cases in real testsets but your model was trained on complete cases only? Subsequently, we have to think about the question: "Are the test and validation datasets complete or incomplete?" An approach known as semi-supervised learning selects merely labeled records for the validation sets but allows unlabeled records to be part of the trainingset (Hastie et al., 2009; James et al., 2014). Although the method might be a useful solution to deal with missing outcome values, it is possible that selection effects downgrade the reliability of analyses. My research will cover these questions.

Based on the observed data alone, it is not possible to differentiate between MAR and MNAR. Molenberghs, Beunckens, Sotto, and Kenward (2008) have shown that "Every missingness not at random model has a missingness at random counterpart with equal fit." Therefore, it is common to perform some sort of sensitivity analysis to test the stability of statistical estimates (Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke, 2015). A relatively easy but reliable way to perform such an analysis is by adding a constant δ to the imputed values. Generally, several values of δ are tested and when the analysis outcomes are stable for different δ, we can be confident that the MAR assumption holds (Van Buuren, 2012). Shahab (2015) developed a method to estimate the amount of δ with a so-called random indicator. By iteratively determining this indicator, the imputation procedure can be adapted while running, and as such result in good imputations even under MNAR. Shahab's method is tested for one type of MNAR missingness and in the situation of normally distributed variables. Continuation of his work is required, especially for datasets with categorical variables or with poisson and exponential distributions. In addition, it would be interesting to know whether it is possible to test H_0: $\hat{\delta} = 0$ against H_a: $\hat{\delta} \neq 0$. If so, we could have a beginning of what might be a test to discriminate between MAR and MNAR.

4.3 Accuracy

The third chapter of my research project will compare measures of prediction accuracy with measures of imputation accuracy and statistical validity. With the availability of amputation procedure `ampute` (Chapter 1) and with increased knowledge about missingness mechanisms in the prediction domain (Chapter 2), I will set up a realistic simulation study. In particular, I will simulate a multivariate dataset, generate realistic forms of missing data, execute common missing data methodologies such as listwise deletion and mean, median and most frequent imputation and evaluate prediction accuracy, imputation accuracy and measures of statistical validity. In earlier studies, Acuña and Rodriguez (2004) showed the performance of missing

data methods in classification tasks by means of prediction accuracy, and Tang and Ishwaran (2016) and Stekhoven and Bühlman (2012) evaluated imputation accuracy for random forest imputation models. One of the few studies that compare prediction accuracy with imputation accuracy is presented by Garcia-Laencina et al. (2010). We will extend their research to regression tasks. In particular, we will focus on fast and easy imputation methods because these methods are already implemented in Python and most commonly used by applied scientists. If it turns out that more complicated missing data methods such as stochastic regression imputation, multiple imputation or machine learning methods should be preferred, I will develop these methods in Python-modules.

4.4 Standardization

The development and implementation of standardized analysis techniques is extremely attractive, especially in the data science framework. An example of this is the `.Pipeline` module from Scikit-learn in Python (Pedregosa et al., 2011). Here, the user specifies a list of possible data transformations and a second list with some machine learning methods. The `.Pipeline` module will then automatically try every combination of data transformation and analysis method, including the k-fold crossvalidation procedure and the desired evaluation metrics. The `.Preprocessing` module of Scikit-learn contains a function called `Imputer` which can be implemented in the same pipeline and automatically generates mean, median or most frequent imputations of the missing values. In `R`, comparable trends emerge such as the development of `tidyverse` (Wickham & RStudio, 2017). Additionally, Van der Loo (2017) recently published the `simputation` package which "offers a number of commonly used single imputation methods, each with a similar and hopefully simple interface."

The extent to which imputation strategies can be 'pipelined' needs examination. There may be circumstances where scientific research could benefit from machine learning techniques and pipline practices. For example, Statistics Netherlands recently started research to investigate whether a standardized use of random forests models would give valid statistical estimates (Park, Pannekoek, & Van der Loo, 2017). Park et al. (2017) describe that random forests do not require a specific modelling or variable selection. As such, this machine learning technique is useful in situations where an imputation model needs to be generalized to other datasets, or when new variables are added to an existing, already imputed, dataset. Shah et al. (2014) describe that random forests are especially useful when interactions between variables need to be taken into account. For which situations do these results from Park et al. (2017) and Shah et al. (2014) hold? What guidelines do we need to standardize missing value treatment? And what are the implications for software implementations?

In particular, the fourth chapter of my research project will show the development of a standardized approach to investigate missing data. In general, it is acknowledged that a proper handling of missing data problems requires careful thinking about: the missingness percentage, missing data patterns, underlying mechanisms, possible predictor matrices and very important, the research or analysis question at hand (Rubin, 1987; Van Buuren, 2012). An `R`-package such as `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) therefore includes functions as `md.pattern` to inspect the missing data patterns. Other packages are created especially to assist with a thorough exploration of the missing values (e.g. `narniar`: Tierney, 2017). However, few researchers know how to actually perform such an exploration. Moreover, it is often hard for researchers to relate the results of such an exploration to their choice of missing data method. Therefore, I will develop an `rmarkdown` and Jupyter notebook that processes an incomplete dataset and returns valuable overviews of the missing data and their meaning. The availability of such a standardized exploration method might help all sorts of researchers to take logical decisions. Remark that Python does not yet contain exploration,

amputation or imputation funections, and all of these need to be developed.

4.5 Fusion

Data fusion is the process of combining two or more datasets before doing an analysis (Leulescu & Agafitei, 2013). The method has many applications, such as matching of non-overlapping surveys or matching of surveys to business or administrative data. Data scientists are quite used to combining multiple source data as part of the feature engineering process (Hastie et al., 2009; James et al., 2014).

The development of evaluation measures of fusion procedures is an ongoing process (e.g. De Waal, 2015). Theoretically, the evaluation of a fusion procedure should be done based on four distributional aspects (i.e. the individual, marginal, conditional and joint distribution, Rässler, 2004), but the absence of a true dataset complicates such evaluation. Therefore, in practice, evaluation measures such as the hit rate are used (Rässler, 2004). However, there are circumstances where the hit rate turns out extremely well but the matching procedure is in fact not sufficient. Simulation studies are needed to more deeply assess the reliability of fusion procedures.

5 Organization

This research project is carried out by me, Rianne Schouten, as a PhD student at Utrecht University under supervision of dr. Gerko Vink and prof. Stef van Buuren. Funding for the first year of this project is provided by DPA Professionals, where I was part of the Data Science Excellence Program and learned how to set up a data science pipeline and use machine learning techniques. The second year I am employed as Developer Data & Analytics at Samen Veilig Midden-Nederland.

In general, the planning for my research project is as follows:

Year 0: Development of R-function `ampute`, write a vignette and write an article draft to present an explanation and extensive test of the method. Send `ampute` article for review. **Amputation 1**

Year 1: Evaluate mechanisms for bivariate normal distribution, write article and send for review. Start simulations to translate missing data methodology to data science use cases and start communicating these results in blogposts. Develop amputation and imputation Python functions for the simulation pipeline in Python. **Robustness 1**

Year 2: Based on the outcomes of the simulations in year 2, determine contents for robustness and accuracy papers. Preferably one paper with an explanation of missingness mechanisms in the prediction domain, and one paper with a comparison of prediction and imputation accuracy. Based on the functions from year 1, generate `rmarkdown` and Jupyter notebook documents. Write blogposts to proof the practical use of these documents and communicate the notebooks with businesses. **Robustness 2, Accuracy 1**

Year 3: See what the field needs. Either start research on the random indicator (RI) method (pillar: robustness) or start research on the evaluation of fusion techniques (pillar: fusion). **Robustness 3 or Fusion 1**

References

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect in the classifier accuracy. In D. Banks, F. McMorris, P. Arabie, & W. Gaul (Eds.), *Classification, clustering and data mining applications* (pp. 639 – 647). Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago.

Brand, J., Van Buuren, S., Groothuis-Oudshoorn, K., & Gelsema, E. (2003). A toolkit in sas for the evaluation of multiple imputation methods. *Statistica Neerlandica*, *57*(1), 36–45.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330 – 351.

De Waal, T. (2015). *Statistical matching: Experimental results and future research questions* (Tech. Rep.). Statistics Netherlands, Discussion paper.

Garcia-Laencina, P., Sancho-Gomez, J., & Figuiras-Vidal, A. (2010). Pattern classification with missing data: a review. *Neural Computations & Applications*, *19*, 263 – 282.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer, New York.

Horton, N., & Kleinman, K. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Journal of the American Statistical Association*, *61*(1), 79 – 90.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning.* Springer, New York.

Leulescu, A., & Agafitei, M. (2013). *Statistical matching: a model based approach for data integration* (Tech. Rep.). Eurostat Methodologies and Working Papers.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B*, *70*, 371–388.

Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., & Verbeke, G. (2015). *Handbook of missing data methodology.* Chapman & Hall/CRC Press: Boca Raton.

Park, S., Pannekoek, P., & Van der Loo, M. (2017). Random forests for official statistics imputation: towards a more efficient methodology.
(Under review)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peeters, M., Zondervan-Zwijnenburg, M., Vink, G., & Van de Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, *12*, 377 – 394.

Rässler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, *33*(2), 153 – 172.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3), 581 – 590.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New York.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147 – 177.

Schouten, R., Lugtig, P., Brand, J., & Vink, G. (2017). Generate missing values with ampute [Computer software manual]. Retrieved from https://rianneschouten.github.io/mice_ampute/vignette/

ampute.html

Schouten, R., & Vink, G. (2017). The dance of the mechanisms: how observed information influences the validity of missingness assumptions.
(Under review)

Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missig data using mice: a caliber study. *American Journal of Epidemiology*, *179*, 764 – 774.

Shahab, J. (2015). Imputation under a nonignorable missingness mechanism. In *Dual imputation strategies for analyzing incomplete data* (pp. 47 – 62). (Dissertation)

Stekhoven, D., & Bühlman, P. (2012). Missforest non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*, 112 – 118.

Tang, F., & Ishwaran, H. (2016). Random forest missing data algorithms. *Statistial Analysis and Data Mining: The ASA Data Science Journal*, *10*, 363-377.

Tierney, N. (2017). Getting started with naniar [Computer software manual]. Retrieved from https://github.com/njtierney/naniar/blob/master/vignettes/getting-started-w-naniar.Rmd

Van Buuren, S. (2012). *Flexible imputation of missing data.* Chapman & Hall/CRC.

Van Buuren, S., & Groothuis-Oudshoorn, C. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3).

Van der Loo, M. (2017). Getting started with simputation [Computer software manual]. Retrieved from https://github.com/markvanderloo/simputation/blob/master/pkg/vignettes

Wickham, H., & RStudio. (2017). Tidyverse [Computer software manual]. Retrieved from https://www.tidyverse.org/