NA-Missing Value Literature Review

Paul A. Hodgetts

17 October 2020

Contents

1 Overview **2**

2 Annotated Bibliography **2**

2.1 Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*(4), 1012-1028. 2

2.2 Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., & Furrer, R. (2018). Predicting missing values in spatio-temporal remote sensing data. *Transactions on Geoscience and Remote Sensing*, *56*(5), 2841-2853. 2

2.3 Gomez-Carracado, M. P., Andrade, J. M., Lopez-Mahia, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, *134*, 23-33 3

2.4 Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1–31 3

2.5 Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). Using sex and gender in survey adjustment. 3

2.6 Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1-10 3

2.7 Templ, M., Alfons, A. & Filzmoser, P. (2012). Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, *6*, 29–47. 4

2.8 Tierney, N. (2017). visdat: Visualising whole data frames. *Journal of Open Source Software*, *2*(16), 355. 4

2.9 Viano, S., & Baker, D. J. (2020). How administrative data collection and analysis can better reflect racial and ethnic identities. *Review of Research in Education*, *44*(1), 301-331 4

2.10 Yadav, M. L. and Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*, 104–118. 4

1 Overview

The following annotated bibliography presents 10 articles in relation to the topic of NA or missing values from various fields of research. They range in field of research from the family studies to earth sciences. Topics covered include the various methodologies of missing value imputation to the importance of understanding the creation of missing values. Recommendations of best practices regarding missing values are discussed from various perspectives. Such as, an understanding of the humans behind the creation of the missing values and the importance of including missing values methods for research transparency.

There are many more articles regarding the topic of missing values, but the general theme of research seemed to regard the topic of imputation over understanding or reproducibility/transparency. That the immediate step to treating missing values is to remove or replace the values is of concern for transparency and reproducibility in research. As well that it presents concerns for survey and social science research regarding the discrimination of those people associated with the missing values. Further research into this topic would provide a better understanding as to current and future practices regarding missing values, as well as the creation of a toolkit that could assist researchers in working with missing values for the purposes of transparency, reproducibility, and understanding of missing values.

2 Annotated Bibliography

2.1 Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*(4), 1012-1028.

In this article, Acock summarises the types of missing values (i.e. missing by definition of the subpopulation, missing completely at random, missing at random, and NI missing values); the traditional approaches to working with missing values (i.e. listwise deletion, pairwise deletion, mean substitution, and indicator/dummy variable adjustment); and more modern approaches to working with missing values (i.e. single imputation using EM, multiple imputation, and full information maximum likelihood approaches). Acock additionally provides empirical examples of the approaches and criticisms as to their uses. Lastly, Acock provides a list of recommendations as to the approaches discussed when working with missing values.

This article provides a good understanding as to approaches to missing values in social sciences, although the prevalence of some traditional approaches may not be as common in 2020 as they were in 15 years ago. Additionally, the recommendations that Acock provides are useful in the practice of working with missing values outside of the practice of deletion and imputation.

2.2 Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G., & Furrer, R. (2018). Predicting missing values in spatio-temporal remote sensing data. *Transactions on Geoscience and Remote Sensing*, *56*(5), 2841-2853.

This article provides insight into another area of research in which missing values are experienced, and provides a novel approach to managing those missing values. Discusses the importance of missing values to earth sceience research and the ways in which missing value may occur. Additionally, discusses within the conclusion of the article the importance of making procedures available through R packages for the sake of transparency. Overall, provides important insights into an area of research outside the social sciences regarding the methodologies used towards missing values.

2.3 Gomez-Carracado, M. P., Andrade, J. M., Lopez-Mahia, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, *134*, 23-33

This article from Gomez-Carracado et al. provides an understanding of missing values outside the realm of social sciences. The article provides an understanding as to the issue of missing values in environmental science and the methods of working with missing values. When discussing missing values it is important to understand how they appear and are understood in various fields of research. There is not a one-size fits all approach to missing values.

2.4 Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, *70*(1), 1–31

This article from Josse and Husson provides an overview of the missMDA R package that performs principal component methods on incomplete data sets. The missMDA package also provides the user with the capability of performing single imputation. This article is important as to understanding what is available in terms of R packages in managing missing values.

2.5 Kennedy, L., Khanna, K., Simpson, D., & Gelman, A. (2020). Using sex and gender in survey adjustment.

An important article that recognizes the importance of missing values in that a missing value does not mean devoid of information. Discusses the importance of understanding that there is a human behind the values within a dataset and that simply removing or imputing the data with a new value is not representative of that human. Provides important insight as to current approaches to missing values within statistical and social research.

2.6 Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology*, *57*(1), 1-10

This article from Scholmer et al. provides an important analysis into the approach to missing values in counseling psychology research as well as providing recommendations as to future procedures for dealing with missing values. These recommendations include that researchers attend better toward missing values and better explain the methodologies used in dealing with missing values. These recommendations are encouraged under the practice of accurate and informed science.

In addition to providing a set of recommended best practices, this article provides further insight into the methodological approach to missing values in a field of research beyond social surveys.

2.7 Templ, M., Alfons, A. & Filzmoser, P. (2012). Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification*, *6*, 29–47.

This article by Templ et al. provides an overview of the VIM R package, which allows R users to graphically explore missing values within a dataset so as to gain a better understanding of the data. While the VIM packages provides important analysis tools for missing values, it does not provide a set of tools that aid in working with missing values according to recommended best practices mentioned in other articles. Additionally, it is important to understand what packages are available to R users and the current approach to the development community regarding missing values.

2.8 Tierney, N. (2017). visdat: Visualising whole data frames. *Journal of Open Source Software*, *2*(16), 355.

A short paper on the visdat package that provides users with the ability to visually explore a dataset. This is similar in regards to the VIM package in that it provides R users with functions that allow the graphical exploration of missing values in a given dataset. Additionally, while such packages provide a better understanding as to the composition of a dataset, the visdat package does not provide tools as to best practices for missing values that would assist in the reproducibility and validity of a study.

2.9 Viano, S., & Baker, D. J. (2020). How administrative data collection and analysis can better reflect racial and ethnic identities. *Review of Research in Education*, *44*(1), 301-331

This article from Viano and Baker provides important insights into the collection of racial and ethnic identities data and the analysis of that data. Presents examples in which the use of data collected in different periods of time can create missing values due to changes in how racial and ethnic identities are presented. Importantly presents how missing values can be created within this field of research and the challenges faced when working with said data. An important piece for understanding human-first over number-first.

2.10 Yadav, M. L. and Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*, 104–118.

This article provides an overview of R packages that provide users with the ability to impute missing values. Provides an important overview of more than one package and further insight as to the approach to missing values within the R development community. That is, missing values are something to be replaced or fixed.