Journal of Statistical Software

July 2018, Volume 85, Book Review 4. *doi: 10.18637/jss.v085.b04*

Reviewer: Hakan Demirtas
University of Illinois at Chicago

Flexible Imputation of Missing Data

Stef van Buuren
Chapman & Hall/CRC, Boca Raton, 2012.
ISBN 978-1439868249. xxv + 316 pp. USD 107.02 (Hardcover).
`http://www.stefvanbuuren.nl/mi/FIMD.html`

Missingness is a commonly occurring phenomenon in many applications. Determining a suitable analytical approach in the absence of complete observations is a major focus of scientific inquiry due to the extra sophistication that arises through missing data. Incompleteness generally complicates the statistical analysis in terms of reduced statistical power, biased parameter estimates, and degraded confidence intervals, and thereby may lead to false inferences. Developments in computational statistics have produced flexible missing-data procedures with a sound statistical basis. One of these procedures involves multiple imputation (MI), which is a stochastic simulation technique in which the missing values are replaced by $m > 1$ simulated versions. Subsequently, each of the simulated complete data sets is analyzed by standard methods, and the results are combined into a single inferential statement that formally incorporates missing-data uncertainty to the modeling process. MI has gained widespread acceptance and popularity in the last few decades. It has some well-accepted advantages: First, MI allows researchers to use conventional models and software; an imputed data set may be analyzed by literally any method that would be appropriate if the data were complete. As computing environments and statistical models grow increasingly complex, the value of using familiar methods and software becomes more pronounced. Second, there are still many classes of problems for which no direct maximum likelihood procedure is available. Even when such a procedure exists, MI can be more attractive due to fact that the separation of the imputation phase from the analysis phase lends greater flexibility to the entire process. Lastly, MI singles out missing data as a source of random variation distinct from ordinary sampling variability.

Van Buuren's work is one of the few books that exclusively focus on MI. The book can be regarded as an extended tutorial on the practical application of the R package **mice** that is available on the CRAN (Comprehensive R Archive Network). The name of the package stands for "multiple imputation by chained equations". This MI technique (also known as fully conditional specification-FCS) is built upon a series of univariate models. Sequential regression equations are specified in a cyclic fashion, by taking the type and nature of a variable given others (e.g., logistic model for a binary variable, proportional odds model for an ordinal variable, linear regression for a normal variable, etc.) into consideration. The set of conditional specifications collectively defines a Gibbs' sampler. A major competitor in the parametric domain to conduct MI is joint modeling (JM) where a joint distribution is posited for all variables in the system. However, formulation of the best possible multivariate distribution that includes all usual variable types may be a daunting task. MI with sequential regression/chained equations lacks theoretical justification and may suffer from problems such as incompatible conditionals, multivariate suboptimality, and feedback loops in the form of dependence of missingness on other missing variables. A sequence of valid conditional distributions may not correspond to a unique joint distribution. There is no reason to engage in fruitless philosophical discussions about JM versus FCS, the goal of MI is modeling the uncertainty due to missing data in addition to the ordinary sampling variability in a reasonable way, so theory obsession is not really necessary. The good news is that FCS works in a wide spectrum of practical settings, and increasing number of people are using FCS to carry out MI. Instead of trying to formulate a difficult-to-defend joint system in the presence of all types of variables, it is more important to incorporate information from other variables and unique characteristics of the data set such as skip patterns, nonlinear relations, and interactions.

The book is primarily geared towards applied people, and the intended audience is biostatisticians, epidemiologists, methodologists, substantive researchers, and practitioners who are not card-carrying statisticians. As the mathematical details are mostly avoided and the main orientation of the book is on the practical side, the required background is not very heavy. Familiarity with statistical methods and multivariate statistics would suffice to grasp the fundamental ideas in the book. The book generates a favorable impression by and large, but it also has some problems. Its merits outweigh its downside. So, let us start with the shorter list. Negatives include:

- The three parts of the book are of very different lengths, they could have been organized in a better way and more evenly constructed.

- One thing that persists throughout the book is the false premise that imputed data and observed data should be similar. We all know that it holds only under MCAR (missing completely at random) mechanisms. Yes, imputed data should feel real, but there is an over-emphasis and over-reliance on this; inexperienced readers might get the wrong idea that the differences between imputed data and observed data should be minimal.

- Some graphs are incomprehensible mostly due to color mix-ups, and there are at least a few text-graph mismatches and inconsistencies.

- Writing is occasionally choppy. There are some suboptimal and/or controversial statements that not everybody would agree on or that are discordant with the established terminology (phrases such as "the best general method", "it is often the case that MI is more efficient than complete case analysis", and "to correct the missing data, to correct for the nonresponse" as if it is something to be corrected, "adjust" would be the right word here. Hard-to-make-sense kind of rough places from a reader's perspective and some typographical errors are noticeable.

On the other side of the coin, here are the positives:

- A broad range of practical advices and guidance are provided.

- Algorithms are given for users who want to go beyond what the package does and really understand how the procedures are applied.
- End-of-chapter exercises are a plus and can be valuable for teaching and learning purposes.
- The literature is adequately cited, citations span a satisfactory range.
- The author maintains a website at http://www.stefvanbuuren.nl/mi/FIMD.html. However, the reference manual is more up-to-date and includes more information.
- The book is based on a well-studied and commonly employed package that implements a well-understood MI technique.
- Limitations and shortcomings are honestly described.
- Suggestions on future research directions are compelling.
- Information on other MI software is made available in the Appendix.
- Writing is generally fine and the book is easy-to-follow, not a lot of back and forth page-turning is needed.
- The book touches nearly all major conceptual and practical aspects of MI as well as many other dimensions of the missing-data world.
- The book could be considered a self-study tool as well as a reference book. It can also be a decent supplementary source to an imputation or more generally a missing-data book that has a more pronounced methodological and statistical flavor.

We are now ready to get into the specifics of the coverage. The monograph starts with the foreword of the founder of MI, Donald Rubin. A list of algorithms is provided in the beginning; a list of figures and tables would have been nice. The book consists of three main parts (basics, case studies, and extensions) that span ten chapters. Technical material that goes beyond the fundamentals is made explicit and it is stated that readers can skip such portions in the first reading. Part 1 consists of Chapters 1–6. Chapter 1 starts with a general account for the problem of missing data and the introduction of the major missingness mechanisms based on Rubin's taxonomy. Subsequently, commonly harnessed ad hoc missing data methods (and their somewhat more principled variants) such as listwise and pairwise deletion, mean imputation, regression imputation and its stochastic version, LOCF and BOCF (last and baseline observation carried forward, respectively), and the indicator method are discussed. A summary of these simple approaches and an overview of associated assumptions are given. The procedure of MI and reasons to use it are described, and the goal of the book is defined along with a section of what the book does not cover such as weighting techniques and likelihood-based approaches. Some decent general advice to minimize the missing data rate is given. Chapter 2 covers a historical overview of MI, mentions how increasingly popular it has been getting in recent years, differentiates unit and item nonresponse, discusses some potential causes of missing data, introduces the notation, re-visits the missingness mechanisms, and explains ignorability and its implications. It then moves to why and when MI works, presents different sources of variation in incomplete data, the scope of the MI models with some technical discussion on MI components such as properness, variance ratios, and degrees of freedom, and touches upon how to combine results into a single inferential summary in MI via scalar and multi-parameter inference. The goal of MI is not accurately predicting the missing values, it is properly reflecting the uncertainty due to missing data, and the book correctly emphasizes this. The chapter ends with evaluation criteria such as bias, coverage, and confidence interval length, arguments that relate to when to use MI and how many imputations are needed.

Chapter 3 focuses on univariate missing data. It is a key chapter for the rest of the book. It starts with imputation based on a regression model, then adds a noise, then adds parameter uncertainty. It discusses predictive mean matching in which drawn values (imputations) come from the real, observed data themselves; it describes imputation under the normal linear model under which four methods (predict, predict+noise, Bayesian MI, bootstrap MI) are mentioned and relevant algorithms are given. Generating incomplete data under MAR (missing at random) mechanism is illustrated. Imputation under non-normal distributions is covered. Many important aspects of MI such as predictive mean matching, imputing categorical, count, and semi-continuous data as well as censored, truncated, rounded data, perfect prediction, classification and regression trees, multilevel data are delineated. Techniques that are designed to handle nonignorably missing data such as pattern-mixture models and selection models are explained along with sensitivity analysis, which is common when missingness is nonignorable. Chapter 4 is concerned with multivariate missing data, missing data patterns, influx and outflux (measures for how variables connect to each other) statistics in multivariate imputation, joint and conditional modeling for continuous and categorical data. As mentioned before, JM and FCS have relative advantages (the former has a better theoretical justification, the latter has flexibility and practicality), and a comparison is made. In the multivariate case, there are some issues that need to be dealt with (pertinent to JM, FCS, or both): (a) the predictors themselves can contain missing values; (b) circular dependence may occur; (c) variables are often of different types; (d) collinearity and empty cells can occur; and (e) the ordering of the rows and columns can be meaningful as in longitudinal data. These potential problems as well as implementation-related paradigms such as convergence, incompatible conditionals, and impossible imputations are discussed in depth.

Chapter 5 is about imputation in practice. An imputation model should account for the process that created the missing data, preserve the relations in the data, and preserve the uncertainty about these relations. An overview of modeling choices is given in a step-by-step format, followed by a discussion on ignorability and the choice of model predictors. Subsequently, practical situations that every imputer faces (how to deal with derived variables, sum scores, interaction terms, compositional data, quadratic relations, convergence diagnostics, model fit versus distributional discrepancies) are addressed. I find the conclusion subsection (5.7) and this chapter in general very useful for readers. Chapter 6 proceeds with what to do with the imputed data. Parameter pooling and inference for normal and non-normal quantities are described. Table 6.1 shows suggested transformations towards normality, which could be a good thing to adopt for the beginners. Statistical tests for MI such as Wald test, likelihood ratio test, χ^2 test are outlined; variable selection techniques and model optimism (identifying truly important factors) issues are also covered in this chapter.

Part 2 encompasses Chapters 7–9, and is entitled “Case studies”. As the title implies, this part is concerned with applications to real data. A common theme in Chapter 7 is “problems with the columns”, meaning how to choose truly important columns (variables) when there are too many of them. Some helpful guidance is provided. Some topics such as sensitivity analyses when nonignorable nonresponse is suspected, causes and consequences of missing data, and delta adjustment are elucidated. Chapter 8 pertains to selection issues, and is mostly dedicated to "problems with the rows", meaning that the dropout mechanism may be selective, leading to a systematic bias between complete and incomplete rows. Chapter 9 covers longitudinal data, long and wide format, time raster imputation, intent-to-treat, broken stick model, shrinkage, and the change score. Although Part 2 has respectable scientific merit, it is not particularly entertaining or riveting from an intellectual hedonism standpoint.

Part 3 is named "Extensions", it could have been named better. It consists of Chapter 10 that starts with a discussion on some dangers, do's and don'ts, and continues with reporting guidelines, imputation of potential outcomes, coarsened data, data fusion, planned missingness, verification bias, measurement error, parallel computing, algorithms for blocks and batches, nested imputation, and distribution-free pooling rules. It is followed by an Appendix that includes other software for MI. In my humble opinion, Chapters 3–5 and 10 are the most beneficial parts of the book.

In summary, I recommend this book to anybody who deals with MI at any level as it presents essential ideas and techniques with illustrative examples, in an engaging way, both on conceptual and applied levels. It contains a healthy dose of material and does a great job of explaining the fundamentals and operational features of MI. It is a handy source and a solid reference for practitioners who routinely impute data. Yes, it seems like an extended tutorial on the R package **mice**, but goes far beyond that. At the end of the day, despite its defects, it is an insightful and horizon-broadening book.

Reviewer:

Hakan Demirtas
Division of Epidemiology and Biostatistics (MC 923)
University of Illinois at Chicago, School of Public Health
1603 West Taylor Street, Room 950
Chicago, IL, 60612-4336, United States of America
E-mail: demirtas@uic.edu
URL: http://www.healthstats.org/members/hdemirtas.html

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
July 2018, Volume 85, Book Review 4 *Published:* 2018-07-02
doi:10.18637/jss.v085.b04