Incomplete Data Analysis

Vanda Inácio

University of Edinburgh

Semester 1, 2020/2021

Multiple imputation

↪ In weeks 3 and 4 we have learned about single imputation techniques.

↪ Specifically, we have seen:

1. Mean or unconditional imputation.
2. Conditional mean/regression imputation.
3. Stochastic regression imputation.
4. Hot deck imputation.

↪ All these methods replace the missing value by *one* imputed value. Stochastic regression imputation was a promising approach. Predictive mean matching (a form of hot deck) is also quite good.

Multiple imputation

- ↪ Single imputation, regardless of the method used, fails to satisfy statistical objectives concerning the validity of resulting inferences based on the filled-in data.

- ↪ Because a single imputed value cannot reflect any of the uncertainty about the true underlying value, analyses that treat imputed values just like observed values systematically underestimate uncertainty.

- ↪ Consequently, imputing a single value for each missing datum and then analysing the filled-in data using standard techniques for complete data will result in, for instance, standard errors estimates that are too small, confidence intervals that fail to attain their nominal coverage (as empirically verified in the simulation study of week 4), etc.

"Imputing one value for a missing datum cannot be correct in general because we do not know what value to impute with certainty (if we did, it would not be missing)"
Donald B. Rubin

Multiple imputation

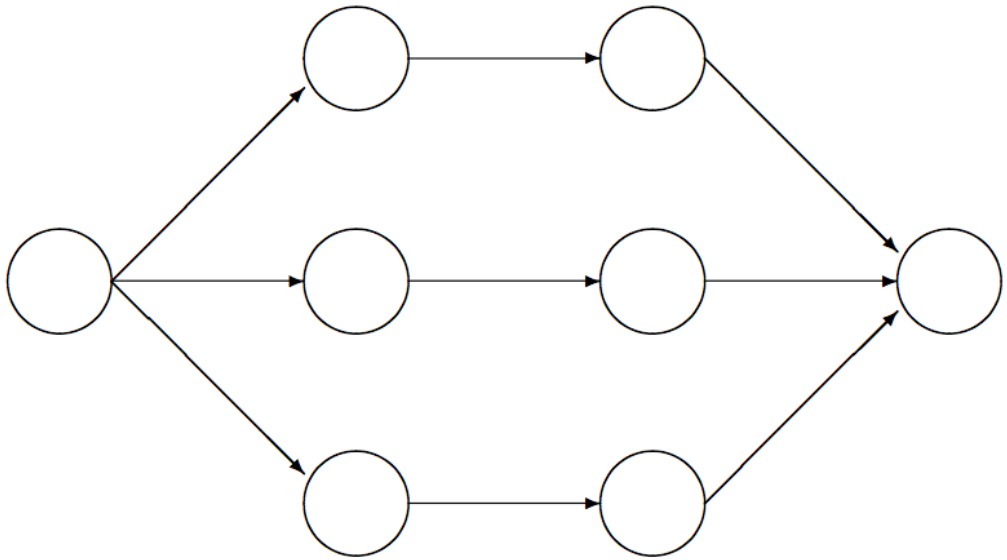
- The problems of single imputation are largely overcome by the use of *multiple imputation* (MI), which is an approach to the missing values problem that allows the investigator to obtain valid assessments of the uncertainty.

- The basic idea of multiple imputation is to impute each missing value several times, thus creating $M > 1$ complete datasets.

- Each of these datasets is analysed by applying the statistical method that we would have used had the data been complete.

- The M results are pooled into a final point estimate plus standard error by simple pooling rules ('Rubin's rules', see next set of slides).

Multiple imputation



Scheme of the main steps in multiple imputation. Here *M* is three. Figure from van Buuren (2012), page 17.

Multiple imputation

- What follows is almost verbatim from van Buuren (2018, pages 30 and 31).
- Multiple imputation was developed by Donald B. Rubin in the 1970's and it is useful to know a bit of its remarkable history.
- The birth of multiple imputation has been documented by Fritz Scheuren (Scheuren, 2005).

Multiple Imputation: How It Began and Continues

Fritz SCHEUREN

1. INTRODUCTION

In the United States, the modern survey sampling revolution began largely at the U.S. Census Bureau. The process, so the story goes, "took off" in the late 1930s when Jerzy Neyman, at the invitation of W. Edwards Deming, came to Washington and lectured at the U.S. Department of Agriculture (USDA 1937; Duncan and Shelton 1978).

The Census Bureau work that led to the two-volume masterpiece by Hansen, Hurwitz, and Madow (HHM, 1953) actually started earlier, as America tried to respond as a country to the Great Depression (Stephan 1949). In any case it is from the perspective of the HHM book that I begin my discussion of the seminal work on imputation that Don Rubin launched in 1977.

Those of you who grew up in another data-collection environment, like experimental design, may object to the narrowness of my focus. After all, other statistical traditions also tackled missingness, albeit in ways different from those used in surveys (e.g., Little and Rubin 2002). The effort to impute, or as it was called then, "allocate" for missingness began in the 1940 Decennial Census and was in full swing by the 1960 Decennial. One of the related factors that contributed to this advance was that the Census Bureau began in the early 1960s making its general-purpose population surveys into public use files (Mulrow and Scheuren 2000).

vented to allocate or fill in the holes in the data matrix. And these were very clever too. The most famous of these was the "Hot Deck" (e.g., Ford 1983).

True to the spirit of that age (and this), technology and statistics were linked. Although computing was primitive, compared to today, there was even a theoretical effort mounted to assess the variance impact of hot deck imputations on the estimates from the pseudo-completed data analyses created, after the data holes had been "filled in" (HHM 1953).

At the very beginning the unit nonresponse problem was handled by duplicating what was operationally a purposive sample of "like" cases to avoid the messiness of not having to deal in the tabulations with the nonresponding cases separately.

CPS response rates were in the high 90s and the hot deck worked well, with little impact on sampling variance. Incidentally the theory developed by HHM assumed, naturally, that random sampling was being used. This gap between theory and practice, it seems safe to say, was probably simply viewed as a minor defect that one could afford to handle in an ad hoc way. In fact, the missingness was so small a problem in these early days that no variance calculations were typically done, although the theory was available—sadly a tradition that continues even today in some Census Bureau surveys, despite the fact that missingness has grown greatly.

Multiple imputation

- Multiple imputation was developed as a solution to a practical problem with missing income data in the March Income Supplement to the Current Population Survey.
- In 1977, Scheuren was working on a joint project of the Social Security Administration and the U.S. Census Bureau.
- The Census Bureau was then using (and still does use) a hot deck imputation procedure.
- Scheuren signaled that the variance could not be properly calculated, and asked Rubin what might be done instead.
- Rubin came up with the idea of using multiple versions of the complete dataset, something he had already explored in the early 1970s.
- According to Scheuren: “The paper is the beginning point of a truly revolutionary change in our thinking on the topic of missingness.”

Multiple imputation

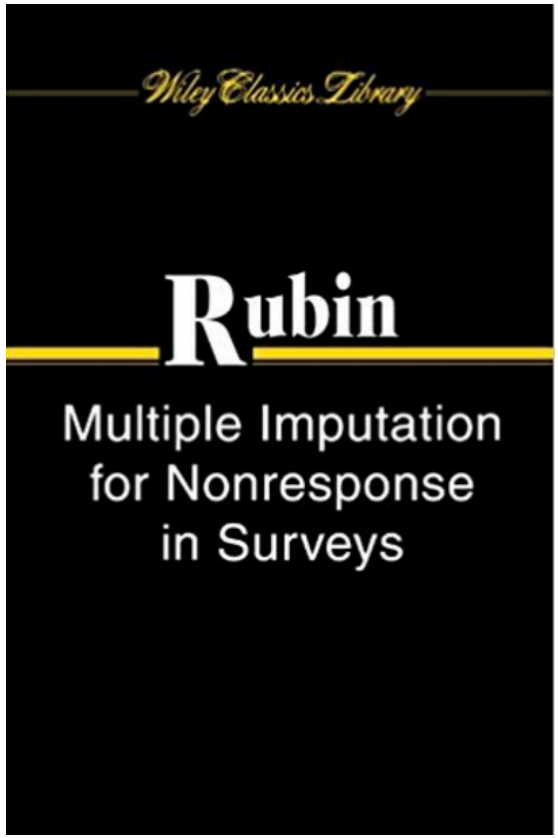
- ↪ Rubin observed that imputing one value (single imputation) for the missing value could not be correct in general.
- ↪ He needed a model to relate the unobserved data to the observed data, and noted that even for a given model the imputed values could not be calculated with certainty.
- ↪ His solution was simple and brilliant: create multiple imputations that reflect the uncertainty of the missing data.
- ↪ The 1977 report explains how to choose the models and how to derive the imputations.

Multiple imputation

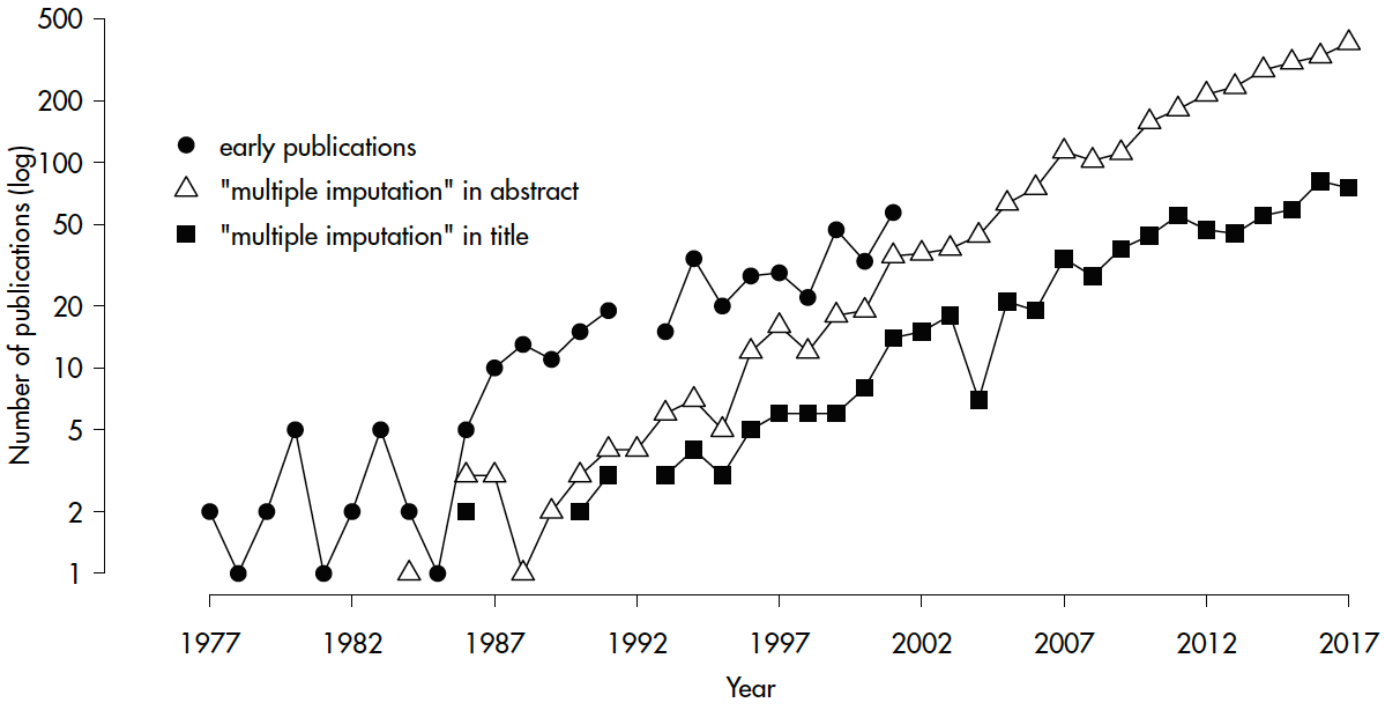
- Over the years, multiple imputation has been sometimes harshly criticized.
- The most severe criticism was voiced by Fay (1992), who pointed out that the validity of multiple imputation can depend on the form of subsequent analysis.
- He produced ”counterexamples" in which multiple imputation systematically understated the true covariance, and concluded that “multiple imputation is inappropriate as a general purpose methodology.”
- Meng (1994) then pointed out that Fay’s imputation models had some flaws and suggested ways to fix it.
- 40 years (more precisely 43 years!) later, multiple imputation is almost universally accepted and it is regarded as the best general method to deal with incomplete data in many fields. In fact, multiple imputation acts as the benchmark against which newer methods are being compared.

Multiple imputation

↪ The first book on multiple imputation (Rubin, 1987).



Multiple imputation



Multiple imputation at the age of 40. Number of publications (log) on multiple imputation during the period 1977–2017 according to three counting methods. Data source: `www.scopus.com` (accessed January 14, 2018). Figure from van Buuren (2018), page 32.

Multiple imputation

- The figure in the previous slide contains three time series with (log) counts on the number of publications on multiple imputation during the period 1977–2017.
- The right most series corresponds to the number of publications per year that featured the search term ‘multiple imputation’ in the title.
- These are often methodological articles in which new adaptations are being developed.
- The series in the middle is the number of publication that featured ‘multiple imputation’ in the title, abstract or key words in Scopus on the same search data. This set includes a growing group of papers that contain applications.

Multiple imputation

- The leftmost series is the number of publications in a collection of early publications.
- This collection covers essentially everything related to multiple imputation from its inception in 1977 up to the year 2001.
- Note that the vertical axis is set in the logarithm. Perhaps the most interesting series is the middle series counting the applications. The pattern is approximately linear, meaning that the number of applications is growing at an exponential rate.

Multiple imputation

↪ In addition to articles and books, high-quality software is now available to ease application of multiple imputation in practice.

R

Amelia by James Honaker, Gary King and Matthew Blackwell creates multiple imputations based on the multivariate normal model. Specialities include overimputation (remove observed values and impute) and time series imputation.
BaBooN by Florian Meinfelder that generates multiple imputations by chained equations. The package specializes in predictive mean matching for categorical data, and in imputation in data fusion situations where many records have the same missing data pattern.
cat by Joseph L. Schafer implements multiple imputation of categorical data according to the log-linear model as described in Chapters 7 and 8 of Schafer (1997).
Hmisc by Frank E. Harrell Jr contains several functions to diagnose, create and analyze multiple imputations. The major imputation functions are transcan() and aregImpute(). These functions can automatically transform the data. The function fit.mult.impute() combines analysis and pooling and can read mids objects created by mice.
kmi by Arthur Allignol performs a Kaplan-Meier multiple imputation, specifically designed to impute missing censoring times.
mi by Andrew Gelman, Jennifer Hill, Yu-Sung Su, Masanao Yajima and Maria Grazia Pittau implements a chained equations approach based on Bayesian regression methods. The software allows detailed examination of the fitted imputation model.
mice by Stef van Buuren and Karin Groothuis-Oudshoorn contributed the chained equations, or MICE algorithm. The package allows for a flexible setup of the imputation model using a predictor matrix and passive imputation.
MImix by Russell Steele, Naisyin Wang and Adrian Raftery implements a special pooling method using a mixture of normal distributions.
mitools by Thomas Lumley provides tools for analyzing and combining results from multiply imputed data.
MissingDataGUI by Xiaoyue Cheng, Dianne Cook, Heike Hofmann provides numeric and graphical summaries for the missing values from both discrete and continuous variables. Removed from CRAN.
missMDA by Francois Husson and Julie Josse contains the function MIPCA() that draws multiple imputations from principal components analysis.
miP by Paul Brix can read imputed data created by Amelia, mi and mice to visualize several aspects of the missing data.
mirf by Yimin Wu, B. Aletta, S. Nonyane and Andrea S. Foulkes provides a function mirf() that create multiple imputations using random forests. Removed from CRAN.
mix by Joseph L. Schafer implements the imputation methods based on the general location model as described in Chapter 9 of Schafer (1997).
norm by Joseph L. Schafer implements multiple imputation based on the multivariate normal model as described in Chapters 5 and 6 of Schafer (1997).
pan by Joseph L. Schafer implements multiple imputation for multivariate panel or clustered data using the linear mixed model.
VIM by Matthias Templ, Andreas Alfons and Alexander Kowarik introduced tools to visualize missing data before imputation. Imputation functions include hotdeck() and irmi(), both loosely based on a chained equations approach.
Zelig by Kosuke Imai, Gary King and Olivia Lau comes with a general zelig() function that supports analysis and pooling of multiply imputed data.
There are many R packages that contain methods for single imputation: arrayImpute, ForImp, imputation, impute, imputeMDR, mtsdi, missForest, robCompositions, rrcovNA, sbgcop, SeqKnn and yaImpute. The functions in these packages typically estimate the missing values in some way, rather than taking random draws.

Multiple imputation

Journal of Statistical Software

December 2011, Volume 45, Issue 1. *http://www.jstatsoft.org/*

State of the Multiple Imputation Software

Recai M. Yucel
University at Albany, SUNY

Abstract

Owing to its practicality as well as strong inferential properties, multiple imputation has been increasingly popular in the analysis of incomplete data. Methods that are not only computationally elegant but also applicable in wide spectrum of statistical incomplete data problems have also been increasingly implemented in a numerous computing environments. Unfortunately, however, the speed of this development has not been replicated in reaching to "sophisticated" users. While the researchers have been quite successful in developing the underlying software, documentation in a style that would be most reachable to the greater scientific society has been lacking. The main goal of this special volume is to close this gap by articles that illustrate these software developments. Here I provide a brief history of multiple imputation and relevant software and highlight the contents of the contributions. Potential directions for the future of the software development is also provided.

Multiple imputation

↪ Below I leave the title and abstract of an interesting article published by Donald Rubin in the *Journal of the American Statistical Association*, volume 91, pages 473–489, in 1996.

Multiple Imputation After 18+ Years

Donald B. RUBIN

Multiple imputation was designed to handle the problem of missing data in public-use data bases where the data-base constructor and the ultimate user are distinct entities. The objective is valid frequency inference for ultimate users who in general have access only to complete-data software and possess limited knowledge of specific reasons and models for nonresponse. For this situation and objective, I believe that multiple imputation by the data-base constructor is the method of choice. This article first provides a description of the assumed context and objectives, and second, reviews the multiple imputation framework and its standard results. These preliminary discussions are especially important because some recent commentaries on multiple imputation have reflected either misunderstandings of the practical objectives of multiple imputation or misunderstandings of fundamental theoretical results. Then, criticisms of multiple imputation are considered, and, finally, comparisons are made to alternative strategies.

KEY WORDS: Confidence validity; Missing data; Nonresponse in surveys; Public-use files; Sample surveys; Superefficient procedures.

Multiple imputation

↪ Another very interesting article that came out in 2018 in Statistical Science (available on Learn).

Statistical Science
2018, Vol. 33, No. 2, 142–159
https://doi.org/10.1214/18-STS644

Multiple Imputation: A Review of Practical and Theoretical Findings

Jared S. Murray

Abstract. Multiple imputation is a straightforward method for handling missing data in a principled fashion. This paper presents an overview of multiple imputation, including important theoretical results and their practical implications for generating and using multiple imputations. A review of strategies for generating imputations follows, including recent developments in flexible joint modeling and sequential regression/chained equations/fully conditional specification approaches. Finally, we compare and contrast different methods for generating imputations on a range of criteria before identifying promising avenues for future research.

Key words and phrases: Missing data, proper imputation, congeniality, chained equations, fully conditional specification, sequential regression multivariate imputation.