A unified platform for missing values methods and workflows

2018-03-29

Applicants

Julie Josse, Ecole Polytechnique, France; julie.josse@polytechnique.edu

Nicholas Tierney, Monash university, Australia; nicholas.tierney@monash.edu

The Problem

Missing values are an unavoidable problem when working with data. They occur for many reasons. For example, individuals choose not to answer survey questions, weather measurement devices fail in wet weather, or participants drop out of a study. Missing values are problematic, as most statistical models and visualisation methods require complete data. This problem is exacerbated as more data from different sources become available, which may not have been designed to be analysed together. Missing values are a problem, so what do we do about it?

There are many ways deal with missing data in an analysis. The most common approach, despite decades of research advocating otherwise, is too toss out cases with missing values. At best, this is inefficient; it wastes information from the partially observed cases. At worst, it results in biased estimates, particularly when the distributions of the missing values are systematically different than the observed values. Fortunately, there are better alternatives to tossing out cases (Schafer and Graham 2002; Little and Rubin 2002; Buuren 2012, Carpenter and Kenward (2012)). Some analysts use model-based approaches, integrating likelihoods or posterior distributions over missing values. Some use imputation, creating single or multiple completed datasets. Some use weighting approaches, appealing to ideas from the design-based literature in survey sampling.

R is able to incorporate so many methods due to its modular packaging system. Currently, there are currently 274 R packages on CRAN that mention missing data or imputation in their DESCRIPTION files. These packages serve many different applications. There are imputation packages such as mice, Amelia, and mi (Gelman and Hill 2011; Honaker, King, and Blackwell 2011; Gelman and Hill 2011). There are providing descriptive statistics and visualisations, like naniar, VIM, and MissingDataGUI (Tierney et al., n.d.; Kowarik and Templ 2016; Cheng, Cook, and Hofmann 2015). There are still many other packages developed to handle complex, heterogeneous (categorical, quantitative, ordinal variables) data of large dimension multi-level data, such as missMDA, and MixedDataImpute (Josse and Husson 2016; Murray and Reiter 2015). There are many missing data packages in R. A problem with so many options is navigating the decision paralysis of choice - which one is right for the case at hand. It is not trivial.

This application aims to valorize current work in missing data by creating an infrastructure steering committee (ISC) working group, which will provide a unified platform that lists and organizes existing packages, articles, tutorials, documentation, and workflows for analyses with missing data. The platform will be easy to extend, and will be well documented, so it can easily incorporate future research in missing values. It will also have a focus on fostering a welcoming community.

We hope that such a tool will enable R to consolidate and strengthen its leadership position on the subject.

The Plan

To create this working group, and this platform, we wish to address the objectives in four parts.

Part one: listing available R packages

We will list available packages for working with missing data, with a brief description. These can be used as a task-view on the subject, as there is currently no task view specific to missing data, aside from sections social sciences (*https://cran.r-project.org/web/views/SocialSciences.html*) and Multivariate Statistics (*https://cran.r-project.org/web/views/Multivariate.html*). We intend to make a call out to other package authors and research in missing data in the R community to form an ISC working group. Those who work with and develop tools and analyses for missing data are spread far around the world. We believe that being a part of a group associated with the R Consortium provides a strong focal point for members to rally under. We are planning on contacting key contributors to missing data to join, for example, one person we intend to contact is Stefan van Buuren, whose website *http://www.stefvanbuuren.nl/mi/Software.html* is also often used as a reference on the topic.

Part two: list articles and related works by theme

Similar to part one with R packages, we will list articles and related works, organising by theme. This can be rather laborious work, to ensure robustness we will also put a call out for authors to submit their articles or works, and reviewers to review their placement on the platform/website.

Part three: Tutorials and workflows.

In part three, we provide peer reviewed tutorials and analysis workflows with a variety of different kinds of missing data. For example, these could focus on exploration and visualization at different stages of the analyses, advantages and drawbacks of methods such as likelihood based approaches, and imputation methods. After imputation or other treatment of the missing data, we will encourage assessing the impact on subsequent analyses and inferences.

Part four: future extensions and beyond

By providing a platform and community to discuss missing data in R, software, and approaches and workflows, we are providing a base from which we can grow. For example, this platform could collect datasets to benchmark imputation methods, which is currently not being done anywhere in the world. By having a community involved in this, we can then have useful discussion on the benchmarks and approaches to multiple imputation, even organize challenges to find the best imputation methods, perhaps in a similar fashion to the M4 forecasting competition (https://robjhyndman.com/hyndsight/m4comp/).

In future work for this working group, we would like to discuss more ambitious work, that currently fall outside the charter of the R Consortium, and would also benefit from community discussion and consensus. For example, considering implementing other special types of missing value other than NA, such as STATAs special missing values; altering messages in base R to encourage other approaches than deleting missing observations by default, or at least indicate the risks, etc.

Expected impact of this iniative. Handling missing values is crucial for data analysis, and an easy to use platform provides immense value to the analysis of many users.

Team

Julie Josse is Professor of Statistics at Ecole Polytechnique in France. Julie was trained as an engineer in statistics in an Agronomy University. Her PhD was given the award "Best PhD in Applied Statistics" by the French Statistical Society. She has specialized in missing data, visualization and the nonparametric analyses of complex data structures. Her work was rewarded by a European Union grant in 2013 to increase her research potential and to spend a year at Stanford University. She has published over 30 articles and written 2 books in applied statistics. Her vocation is to push methodological innovation to bring useful application of her research to users, in particular in bio- and food science. Julie Josse has developed packages to transfer her works such as missMDA dedicated to missing values. Her experience on dealing with incomplete data is recognized by the community: she has just compiled a "Statistical Science" special issue to have a snapshot of the state of the art, she organized the first conference on missing value, "MissData" in 2015 and she is often invited to give lectures around the world to share her experience. She is deeply involved in the R community and is part of Rforwards to widen the participation of minorities in the communities.

Nicholas Tierney was trained in Psychology before changing to statistics, recently completing his PhD in statistics at the end of 2018. Nick is currently a research fellow in the department of econometrics and business statistics at Monash University, and a member of the rOpenSci community, and a software carpentary instructor trainer. Nick has written two R packages that focus on data visualisation and approaches for exploring missing data: visdat, and naniar. He has also authored a peer reviewed paper on model based approaches for understanding structure in missing data. In addition to his experience with statistics and R package development, Nick brings experience on wrangling groups of people together to build projects and communities, having been the lead organiser for the first two Australian rOpenSci ozunconferences (2017 and 2018), being an active member of the Statistics Society of Australia (Young Stats rep for Queensland 2013-2016, General council member for Victoria 2017-), and also having been the organisational chair for the Bayesian Research and Applications Group and the NUMBAT. Nick has presented his research using R domestically at local conferences and meetups, and internationally, at UseR in 2017, and also at the Bay Area R User Group, and at Genentech, California.

Research Assistant (RA). The RA will have experience using R, will have some experience with bookdown, blogdown, and IT skills such as webservices (setting up and maintaining a new domain name, basic HTML + CSS). The RA will ideally have research experience (understand how to conduct a literature review and summarise and synthesize literature). It will not be critical for the RA to have experience working with missing data or with R package development, but this would be a plus. The most important skills would be for the RA to have the capacity to learn about new research areas and critically reason with the literature.

We will also contact contributors to missing data, such as François Husson, (Professor of Statistics at Agrocampus Ouest, France), Stefan van Buuren, and Mark van der loo. We also welcome and encourage interest from other R users, developers, and analysts.

Project Milestones

We would like to conduct the project over the following time periods (M1 refers to month 1).

Milestone one: Create guidelines, infrastructure, and call for support. $2K

1. (M1-M4).

- Find suitable research assistant to help with this project.
- Make a call out to members of the R community interested in joining an ISC working group on missing data.
- Establish a Missing Values Task View along with guidelines and principles to decide what is added to this new task view, which may involve giving badges to packages, say for example if it is on CRAN, or if it has undergone a peer review for the package code. This would mean that packages with peer review or other positive flairs (on CRAN, for example), would be easier to search for and "rated higher" than others.
- Create website infrastructure so that it is easy to navigate and to update, likely done with bookdown or blogdown. Note that the source code for this will be hosted publicly on github.

Milestone two: list articles and related works by theme. $4k

2. (M5-M7). List, organize the packages/articles available.

- List all R packages that work with missing data
- Using criteria and guidelines from milestone one, establish high quality R packages and provide more information about these.

- Add flairs and themes to packages
- Similar to above, but for key articles on missing data
- Key to success here is enlisting the help of the members of the community, to submit their articles or works to be reviewed for placement on the platform/website.

Milestone three: Tutorials and workflows. $4K

3. (M8-M11).

- Gather and write documentation and tutorials to better reproduce existing studies and help the users with their own analyses.
- The idea here being to suggest good practices and pipelines to analyse data with missing values, which might involve video.
- Ideally here Nick and Julie would meet face to face.

Milestone four (future work): future extensions and beyond $10K - 20K

4. (M12-M24).

At this point we would like to assess our progress so far on the project and consider some more ambitious goals. To invest more time into this we will require additional funds to continue this project.

- Invest more time in establishing peer review practices for missing data packages. To this end, we would be able to hire a part time research software engineer to assist with an "editing" role.
- Discuss and implement:
- collecting datasets to benchmark imputation methods,
- organize events like a challenge to provide the best imputation methods, in a similar fashion to the M4 forecasting competition (https://robjhyndman.com/hyndsight/m4comp/).
- more universal approaches to multiple imputation that can incorporate many different styles and workflows.
- Organize data benchmark repository.
- Discuss and arrive at a community consensus for features such as special types of missing value (other than NA like STATA/SPSS/SAS special missing values), and providing messages/warnings in base R when missing values are omitted (e.g., in `plot`, `lm` and `table`) that indicate the risk and encourage other approaches than deletion of missings.

How ISC can help

Employing research assistant(s)

Julie and Nick will jointly supervise a research assistant, who will help with the Milestones described.

Meeting with Partners

It will be of great benefit to for both of us to work together face to face for one week or so. In this case it would either be a trip from France to Australia, or Australia to France. We'd like to request support for helping with the long haul flight.

Online materials

These costs will cover a domain name and server costs for us to set up an online community.

Dissemination

The platform we create will serve as the primary dissemination of the project. We will also write blog posts for the R Consortium site, and Nick will also blog about the experience on his personal blog.

we can maybe speak about using social media, twitter, etc… For the team, say that we are going to contact maybe van der loo, and others?

References

Buuren, Stef van. 2012. *Flexible Imputation of Missing Data.* CRC Press.

Carpenter, James, and Michael Kenward. 2012. *Multiple Imputation and Its Application.* John Wiley & Sons.

Cheng, Xiaoyue, Dianne Cook, and Heike Hofmann. 2015. "Visually Exploring Missing Values in Multivariable Data Using a Graphical User Interface." *Journal of Statistical Software* 68 (1): 1–23.

Gelman, Andrew, and Jennifer Hill. 2011. "Opening Windows to the Black Box." *Journal of Statistical Software* 40.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1–47. http://www.jstatsoft.org/v45/i07/.

Josse, Julie, and François Husson. 2016. "missMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software* 70 (1): 1–31. doi:10.18637/jss.v070.i01.

Kowarik, Alexander, and Matthias Templ. 2016. "Imputation with the R Package VIM." *Journal of Statistical Software* 74 (7): 1–16. doi:10.18637/jss.v074.i07.

Little, Roderick J A, and Donald B Rubin. 2002. *Statistical Analysis with Missing Data.* 2nd ed. New York ; Chichester: Wiley.

Murray, Jared S, and Jerome P Reiter. 2015. "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence." http://arxiv.org/abs/1410.0438.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 (2): 147–77.

Tierney, Nicholas, Di Cook, Miles McBain, and Colin Fay. n.d. *Naniar: Data Structures, Summaries, and Visualisations for Missing Data.* https://github.com/njtierney/naniar.