Further ideas for exploring missing data

Nicholas Tierney

Infectious Disease Ecology and Modelling group

Telethon Kids Institute, Perth, WA, Australia

# Introduction

When you do data analysis, you come across missing data. Because I felt so frustrated by how hard it was to handle and wrangle missing data, I wanted to make it easier. In my endeavours I have written two R packages, visdat and naniar, for exploring missing data, and several papers on the topic.

The present goal is to share some ideas on exploring missing data, using naniar, and visdat. To that end, we will focus on four questions.

1. How do we start looking at missing data?
2. How do we explore missingness in variables?
3. How do we explore missingness relationships?
4. How do we explore imputed values?

Questions 1 and 2 were explored in the previous Software Corner article. This article considers questions 3 and 4.

But first, let’s reacquaint ourselves with the data.

# The data

The data used for this paper is measurements of rodents in Kansas, from Hope (2023). The use of this data is inspired by Allison Horst’s “Exploring missing values in naniar” [shiny application](https://allisonhorst.shinyapps.io/missingexplorer/). In this paper we use a different, larger set of the data. For information on the metadata of the paper see [here](http://lter.konza.ksu.edu/content/csm08-small-mammal-host-parasite-sampling-data-16-linear-trapping-transects-located-8-lter). The data set provides various biometric length and weight measurements, for four species of rodents: the Eastern woodrat, Prairie vole, Western harvest mouse, and Deer mouse.

# How to explore missingness relationships?

We can identify key missing variables using vis\_miss(), gg\_miss\_var(), and gg\_miss\_upset(), but for further exploration, we need to explore the relationship amongst the variables in this data: date, species, total\_length, tail\_length, hind\_foot\_length, ear\_length, weight, sex, and age.

## Exploring using bivariate plots

Let’s say that we want to explore the relationship between tail length and ear length. [Figure 1](#fig-example-geom-point) shows a scatter plot of tail length and ear length.

library(tidyverse)  
library(visdat)  
library(naniar)  
ggplot(rodents,   
 aes(x = ear\_length,   
 y = tail\_length)) +   
 geom\_point()

Warning: Removed 530 rows containing missing values or values outside the scale range  
(`geom\_point()`).

|  |
| --- |
| Figure 1: Plot of ear length against tail length. Ear length is on the X axis and tail length is on the Y axis. We learn that there is a reasonable positive correlation of tail length and ear length. |

The problem with this is ggplot removes the missing values. It’s great that we get a warning message but this makes them hard to explore. We can impute missings with values 10% lower than the minimum value in that variable, which puts these values in a margin area on the graphic. This method comes from [ggobi](https://en.wikipedia.org/wiki/GGobi) (Cook and Swayne 2007), and [manet](http://www.rosuda.org/MANET/) (Unwin et al. 1996).

This imputation is wrapped up in the geom\_miss\_point() ggplot2 geom. [Figure 2](#fig-geom-miss-point) illustrates this by exploring the relationship between tail length and ear length from the rodents dataset.

ggplot(rodents,   
 aes(x = ear\_length,   
 y = tail\_length)) +   
 geom\_miss\_point() +   
 scale\_colour\_brewer(palette = "Dark2")

|  |
| --- |
| Figure 2: Improved plot of tail length against ear length, we can now see the missing values are imputed 10% below the minimum value. The green dots on the Y axis represent tail\_length values that have missing ear\_length. There aren’t any missing values on the X axis, because there aren’t times where tail length is missing when ear length is missing. The row of dots in the bottom left corner are missing for both tail length and ear length |

Being a proper ggplot geom, it supports all of the standard features of ggplot2, such as **facets** and **themes** as shown in [Figure 3](#fig-ggmissing-facet).

ggplot(rodents,   
 aes(x = ear\_length,   
 y = tail\_length)) +   
 geom\_miss\_point() +   
 facet\_wrap(~species) +   
 theme\_minimal()

|  |
| --- |
| Figure 3: A faceted version of the improved tail length against ear length plot where each species is split out into its own subplot. |

## Exploring using modelling

The previous article introduced upset plots, as shown in [Figure 4](#fig-gg-miss-upset), to help identify structure in the missigness. We can perform some basic clustering on the missingness and then and learn which variables and their values predict these missingness groups using decision trees (Tierney et al. 2015; Barnett et al. 2017). We start by adding missingness clusters, choosing four based on [Figure 4](#fig-gg-miss-upset).

gg\_miss\_upset(rodents)

|  |
| --- |
| Figure 4: An upset plot of 7 sets of missingness in the rodents data displaying a more nuanced depiction of the patterns of missingness in the data. |

We encourage exploring different numbers of clusters. We can then confirm this pattern using visualisations as shown in [Figure 5](#fig-gg-miss-var-cluster).

rodents\_miss\_clust <- rodents %>%   
 add\_miss\_cluster(n\_clusters = 4)  
gg\_miss\_var(rodents\_miss\_clust, facet = miss\_cluster)

|  |
| --- |
| Figure 5: Number of missings for each variable for each cluster. We see clear patterns emerge where there are two variables missing in cluster two, six variables missing in cluster three, and not as many missings in clusters one and four. |

We use the R package rpart (Therneau and Atkinson 2023) to fit a classification and regression tree (CART) to the data using all variables.

library(rpart)  
rodent\_miss\_cart <- rpart(  
 factor(miss\_cluster) ~ .,   
 data = rodents\_miss\_clust)

Variable importance scores ([Table 1](#tbl-var-imp)) reveal the most important variables for predicting missingness cluster are date and sex.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Variable importance scores for predicting missingness cluster. The most important variables are date and sex.   | variable | importance | | --- | --- | | date | 221.39 | | sex | 167.96 | | hind\_foot\_length | 20.28 | | age | 15.08 | | species | 10.67 | | tail\_length | 7.37 | | weight | 5.18 | |

To better understand these relationships, we would recommend exploring using partial dependence plots using packages such as vip (Greenwell and Boehmke 2020) and other decision tree plots using rpart.plot (Milborrow 2022).

# How do we explore imputed values?

The [simputation](https://cran.r-project.org/package=simputation) package provides a nice interface to imputation. We will impute values for tail\_length and weight using the impute\_lm() function, then visualise the data, as seen in [Figure 6](#fig-simpute-invisible).

library(simputation)  
rodents %>%  
 impute\_lm(tail\_length ~ species + age) %>%  
 impute\_lm(weight ~ species + age) %>%  
 ggplot(aes(x = weight,  
 y = tail\_length)) +   
 geom\_point()

Warning: Removed 180 rows containing missing values or values outside the scale range  
(`geom\_point()`).

|  |
| --- |
| Figure 6: Imputed values are not visible. A plot of tail length by weight. The Imputed tail length values are not visible because we have no way to identify them in the data. |

In [Figure 6](#fig-simpute-invisible) we don’t get any warnings regarding missing observations - because they are all imputed! However this comes at a cost: we don’t know where the imputations are - they are now sort of invisible.

We can track a copy of the missing data locations by using the function nabular(), which binds another dataset to the current one which notes the locations of the missing data. “Nabular” data is a really important idea in naniar, but to keep it brief, for each column with missing values, a new column is created to help identify misingness. For example, a new column called ear\_length\_NA is created:

nabular(rodents) |>   
 select(starts\_with("ear\_length")) |>   
 head()

# A tibble: 6 × 2  
 ear\_length ear\_length\_NA  
 <dbl> <fct>   
1 39 !NA   
2 18 !NA   
3 17 !NA   
4 21 !NA   
5 19 !NA   
6 19 !NA

The key takeaway here is there is now a copy of the data bound to it, with each column ending in \_NA, and the values either being “NA” for missing, or “!NA” for not missing. For more details on the ideas underlying this, and the benefits, we recommend reading our paper, “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations” (Tierney and Cook 2023).

Using the shadow matrix to keep track of where the missings are, you can actually keep track of the imputations, colouring by what was previously missing in tail\_length. For example, let’s create the nabular data, then impute the data for both tail\_length and weight using a random forest, and plot it in [Figure 7](#fig-simpute-visible-lm).

rodents\_nabular <- rodents %>%  
 nabular() %>%   
 as.data.frame()  
  
rodents\_lm\_tail\_imputed <- rodents\_nabular %>%   
 impute\_lm(tail\_length ~ species + date) %>%   
 impute\_lm(weight ~ tail\_length + species + date)  
  
ggplot(rodents\_lm\_tail\_imputed,  
 aes(x = weight,  
 y = tail\_length,  
 colour = tail\_length\_NA)) +   
 geom\_point(alpha = 0.5) +  
 scale\_colour\_brewer(palette = "Dark2")

|  |
| --- |
| Figure 7: Linear model imputed values of tail length in a scatterplot of tail length vs weight. Weight is on the X axis and tail\_length is on the Y axis, and the points are coloured by whether their tail length values are imputed - ‘NA’ indicates a previously missing value that has been imputed. We can see where imputations are added, but they are very concentrated. |

The simputation package has a nice option to add residual noise to the imputations - in this case we can add some normal noise to the observations, where the residuals are draws with replacement from the model residuals. This gives us much greater variation in the imputations.

The different imputation methods are visualised side by side in [Figure 8](#fig-imputed-comparison). We have also included mean imputation, as a naive comparison. This first imputes the data using the residual method, then rowbinds the two datasets together, creating a column called “imputation\_type”, which records which type of imputation was used, either “add\_residual” or “no\_residual”.

rodents\_lm\_tail\_imputed\_res <- rodents\_nabular %>%   
 impute\_lm(tail\_length ~ species + date,   
 add\_residual = "observed") %>%   
 impute\_lm(weight ~ tail\_length + species + date,  
 add\_residual = "observed")  
  
rodents\_mean\_imputed <- rodents %>%  
 nabular() %>%   
 as.data.frame() %>%   
 impute\_mean\_all()  
  
rodents\_imputed\_comparison <- bind\_rows(  
 add\_residual = rodents\_lm\_tail\_imputed\_res,  
 no\_residual = rodents\_lm\_tail\_imputed,  
 mean\_imp = rodents\_mean\_imputed,  
 .id = "imputation\_type") %>%   
 as\_tibble()

ggplot(rodents\_imputed\_comparison,  
 aes(x = weight,  
 y = tail\_length,  
 colour = tail\_length\_NA)) +   
 geom\_point(alpha = 0.5) +   
 scale\_colour\_brewer(palette = "Dark2") +   
 facet\_wrap(~imputation\_type)

|  |
| --- |
| Figure 8: Comparing imputation methods of tail length in a scatterplot of tail length vs weight. Weight is on the X axis and tail\_length is on the Y axis, and the points are coloured by whether they are imputed - ‘NA’ indicates a previously missing value that has been imputed. We learn that mean imputation provides imputations that are not representative, linear models with no residuals are representative of the data but very concentrated, and adding residuals adds much more variation to your data. |

# Conclusion

In these two software corner articles we have demonstrated the use of the visdat and naniar R packages for exploring and understanding missing data. To find out more please take a look at the vignettes for [visdat](https://cran.r-project.org/web/packages/visdat/vignettes/using_visdat.html) and [naniar](https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html).

# References

Barnett, Adrian G., Paul McElwee, Andrea Nathan, Nicola W. Burton, and Gavin Turrell. 2017. “Identifying Patterns of Item Missing Survey Data Using Latent Groups: An Observational Study.” *BMJ Open* 7 (10): e017284. <https://doi.org/10.1136/bmjopen-2017-017284>.

Cook, Dianne, and Deborah F. Swayne. 2007. *Interactive and Dynamic Graphics for Data Analysis: With and* . Use R! New York: Springer. <https://doi.org/10.1007/978-0-387-71762-3>.

Greenwell, Brandon M., and Bradley C. Boehmke. 2020. “Variable Importance Plots—an Introduction to the Vip Package.” *The R Journal* 12 (1): 343–66. <https://doi.org/10.32614/RJ-2020-013>.

Hope, Andrew. 2023. “CSM08 Small Mammal Host-Parasite Sampling Data for 16 Linear Trapping Transects Located in 8 LTER Burn Treatment Watersheds at Konza Prairie.” Environmental Data Initiative. <https://doi.org/10.6073/PASTA/F7BFB7226B093C817DC391C34A88B514>.

Milborrow, Stephen. 2022. *Rpart.plot: Plot ’Rpart’ Models: An Enhanced Version of ’Plot.rpart’*. <https://CRAN.R-project.org/package=rpart.plot>.

Therneau, Terry, and Beth Atkinson. 2023. *Rpart: Recursive Partitioning and Regression Trees*. <https://CRAN.R-project.org/package=rpart>.

Tierney, Nicholas, and Dianne Cook. 2023. “Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations.” *Journal of Statistical Software* 105 (7): 1–31. <https://doi.org/10.18637/jss.v105.i07>.

Tierney, Nicholas, Fiona A. Harden, Maurice J. Harden, and Kerrie L. Mengersen. 2015. “Using Decision Trees to Understand Structure in Missing Data.” *BMJ Open* 5 (6): e007450. <https://doi.org/10.1136/bmjopen-2014-007450>.

Unwin, Antony, George Hawkins, Heike Hofmann, and Bernd Siegl. 1996. “Interactive Graphics for Data Sets with Missing Values—MANET.” *Journal of Computational and Graphical Statistics* 5 (2): 113–22. <https://doi.org/10.1080/10618600.1996.10474700>.