# Visual Analysis of Missing Data – To See What Isn't There

Sara Johansson Fernstad and Robert C Glen *
University of Cambridge

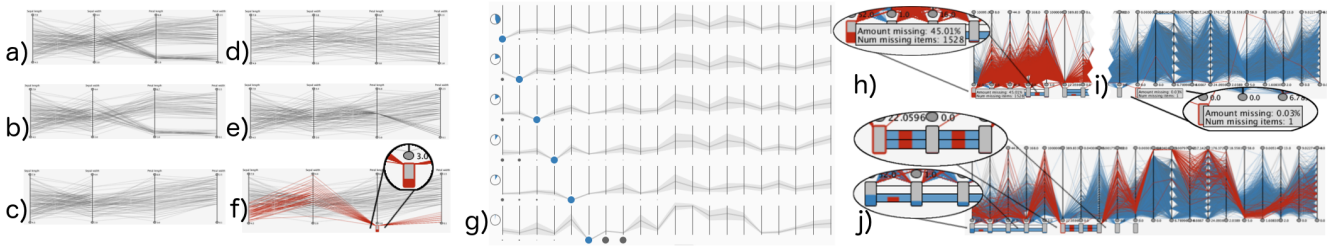


Figure 1: a-c) Missing data examples using the Iris dataset: a) full dataset; b) 33% of items in third variable uniformly missing; c) lowest 33% of items in third variable missing. d-f) The effect of analysis methods on the dataset in figure (c): d) Removal of items with missing values. e) Imputation of mean value. f) Visualization of missing values, highlighting items with missing in red. g) Profile plots of data from a study of skill learning [7], displaying the multivariate missingness profiles of the first six variables. h-j) Examples of missingness patterns using the skill learning study data [7]: h-i) Highlighting in red items with missing values in a variable with high AM, 45%, in (h) and with low AM, 0.03%, in (i). Bars below axes represent the relative amount missing in each variable. j) items with missing values in sixth variable are highlighted in red. The broad red band between missings in the sixth, seventh and eighth variables indicate high JM with the sixth variable. High values in, for instance, the eleventh variable suggests that the probability of missingness in the sixth variable may be conditional upon values in the eleventh variable.

## Abstract

Missing data are records that are absent from a data set. They are data that were intended to be recorded, but for some reason were not. Missing values are common in data analysis and occur in almost any domain, causing problems such as biased results and reduced statistical rigour. Visual analytics has great potential to provide invaluable support for the investigation of missing data. This poster aims to highlight the importance of analysing missing data and the challenges involved, as well as to emphasize the lack of visualization support in the area and through this encourage visualization scientists to discuss and address this highly relevant issue.

**Keywords:** Missing data, visualization, exploratory analysis.

**Index Terms:** I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques; I.5.m [Computing Methodologies]: Pattern Recognition—Miscellaneous

## 1 Introduction

In the context of data analysis, missing data refers to data records that for some reason have not been obtained. Missing records can be thought of as potentially important information that is not available. Since the record is missing, it is unknown to the analyst what value it would have taken had it been observed and, at best, a reasonably good estimate of a probable value can be obtained. Missingness introduces elements of uncertainty and may heavily bias results, causing problems such as loss of statistical power, less reliable estimates and distortion of statistical properties. The development of computational methods for dealing with missing data is a research topic in its own right and the selection of appropriate methods requires understanding of the patterns of missingness. Most approaches assume that missing values are errors that need to be dealt with. However, the fact that values are missing may carry potentially valuable information. By treating missing data as information bearing signals the investigation of missingness may generate novel and valuable insights. This investigation would be greatly facilitated by tools supporting visual analysis of missing records. Nonetheless, the topic of missing data has to a great extent been overlooked by the visualization society. With this poster we aim to highlight the prevalence of missing data and the challenges and opportunities it generates; suggesting visual representations and tasks of particular relevance for visual analysis of missing data.

*e-mail: sf510, rcg28@cam.ac.uk

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

## 2 Why Are Data Missing?

Missing data exist in almost any domain and may have a great variety of causes [1]. In studies such as demographic or consumer surveys, respondents may avoid answering particular questions. In longitudinal studies some participants may not take part in all steps. Subjects may drop out or be excluded from clinical trials, for instance due to response to a treatment. In laboratory based studies physical properties may be unrecordable for certain samples. Values may also be missing due to processing issues or technical limitations. Using data from multiple sources, missingness may be caused by mismatches between databases or variations in naming conventions.

## 3 Analysis of Data With Missing Values

Analysis of data with missing values can be complex. Variables with missing values may be thought of as multi-type variables; the recorded values are of one type and the missing values are a type of their own, which is neither numerical nor categorical. The missing values have no mean or distribution and it is not possible to apply standard statistics to them. Moreover, the degree of missingness and distribution of missing values may affect the appropriateness of analysis methods. Missing values cause uncertainty, but in contrast to many uncertainty problems issues with missing data are not normally overcome by increased sample size. The number of missing values may often increase along with the data, and for high dimensional data there is a high probability that most samples have at least one value missing. The effect of missing values depends both

on the missingness patterns and on how the missing values are handled. Figure 1(a-f) displays examples of this using the Iris dataset. The full dataset without missing values is displayed in figure 1(a), in 1(b) 33% of the *PetalLength* records (third axis) are uniformly missing, and in 1(c) the lowest 33% of the *PetalLength* records are missing. The *PetalLength* mean value in the figures are 3.1, 3.7 and 4.9 respectively. The two main approaches for dealing with missing values are removal and imputation [1]. Removal is when data points or items with missing values are removed prior to analysis, as in figure 1(d). Imputation is when missing values are replaced by estimates, as in figure 1(e). There are many imputation methods available, ranging from simple replacement with mean to complex multiple imputation models. Depending on which method is used, removal and imputation introduce a risk of biased analysis and may lead to too small variability.

## 4 Visual Exploration of Missingness

Many methods for dealing with missing data imply that missing values are errors that have to be removed or replaced. However, the fact that values are missing may carry valuable information, independent of what values they would have taken if they had been recorded. Exploration of missingness patterns may not only support the selection of appropriate methods for dealing with the missing values, but may also provide additional understanding. Important insights could be gained by treating the missing data records as information bearing signals, rather than getting rid of them, and visual analytics can bring valuable contributions to this analysis. By visualizing the missing values ("the data that isn't there"), visual analytics enables exploration of missingness in ways other analysis approaches cannot. Nevertheless, only a small number of publications focus on visual analysis of missingness [6, 4, 9, 5] and a recent paper by Wong and Varga [8] identifies the problem of representing missing data as a particularly important issue for visual analytics systems to address. The value of displaying missing data was also emphasized in a recent text by Kirk [3]. Eaton et al. [2] evaluate approaches to displaying missing data, concluding that a poor indication of missingness has a clear negative effect on interpretation and suggest that visual representations should be enhanced by attributes indicating the existence of missing values. Building on this, figure 1(f) displays a suggestion of how missing values may be represented in parallel coordinates. The plot is enhanced with a 'missingness indicator' below each axis that contains missing values. The size of the indicators coloured bar corresponds to the amount missing in the variable. Interactive highlighting of items with missing values enables exploration of patterns relating to missingness. In the figure, items highlighted in red for instance indicate a relationship between missing values in the third variable and low values in the fourth variable. Figure 1(j) displays an additional example where items with values missing in both of two adjacent axes are represented by a blue band. The height of the band represents the deviation from the expected number of missing values, based on the number of missing values in the two variables, and the red band in the centre represents the currently selected missing values.

We also introduce missingness profile plots (figure 1(g)) to provide an overview of multivariate missingness patterns. A public dataset from a video game skill learning study [7] is used, which displays various patterns relating to missing values. Each profile plot displays the means and standard deviations of items that are missing in a focus variable. Every plot represent one focus variable, highlighted by a blue dot below its axis. The pie chart to the left displays the relative amount missing within the focus variable. Grey dots below non-focus axes represent the amount of items concurrently missing in the focus variable and corresponding non-focus variable. The grey dots below the seventh and eighth axis in the bottom plot indicate that most items missing in the focus variable (sixth axis) are also missing in the seventh and eighth variable.

### 4.1 Patterns of Interest

An important part of exploring complex data is the identification of interesting patterns. This can be supported by the use of quality metrics (QM), which measures the level of 'interestingness' in data. QM is a popular visual analytics concept particularly in the context of large and high dimensional data. Many metrics focus on clustering, classification and correlation but, due to the characteristics of missing data, they may not be appropriate descriptors for missingness patterns. We have, through literature review and discussions with practitioners, defined three main patterns of relevance for analysis of missingness and suggest these as basis for future QM design. Figures 1(h-j) display examples of these patterns.

- **Amount Missing (AM)**: Provide understanding of the distribution of missing values across variables (figure 1(h-j)).
- **Joint Missingness (JM)**: Identification of sets of variables with high joint probability of missing values (figure 1(j)).
- **Conditional Missingness (CM)**: Identification of variables where the probability of missingness is conditional upon recorded values (figure 1(j)).

## 5 Conclusion and Future Work

Missing data is common in many domains and computational methods for dealing with missing data is a research topic in its own right. Few visualization methods support investigation of missingness. We suggest that new insights may be gained by treating missing data as information bearing signals and through this support visual exploration of missingness patterns. This poster aims to encourage further investigation into the potential of visual analysis of missingness. Future work may, for instance, include the development of methods for exploration of missingness patterns in large and high dimensional data; the design of appropriate quality metrics to further support gaining of insights from missingness, as well as additional research into visual representations of missing records.


## Acknowledgements

This work was funded by Unilever Discover Port Sunlight, UK.


## References


[1] J. Carpenter and M. Kenward. *Multiple Imputation and its Application*. Wiley, 2013.

[2] C. Eaton, C. Plaisant, and T. Drizd. Visualizing missing data: graph interpretation user study. In *Human-Computer Interaction-INTERACT 2005*, pages 861–872. Springer, 2005.

[3] A. Kirk. Visualizing zero: How to show something with nothing. http://blogs.hbr.org/2014/05/visualizing-zero-how-to-show-something-with-nothing/, May 2014.

[4] D. F. Swayne and A. Buja. Missing data in interactive high-dimensional data visualization. *Computational Statistics*, 13:15–26, 1998.

[5] M. Templ and P. Filzmoser. Visualization of missing values using the R-package VIM. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology*, 2008.

[6] M. Theus, H. Hofmann, B. Siegl, and A. Unwin. Manet extensions to interactive statistical graphics for missing values. In *In New Techniques and Technologies for Statistics II*, pages 247–259. IOS Press, 1997.

[7] J. Thompson, M. Blair, L. Chen, and A. Henrey. Video game telemetry as a critical tool in the study of complex skill learning. *PLoS ONE*, 8(9), 2013.

[8] B. L. W. Wong and M. Varga. Black holes, keyholes and brown worms: Challenges in sense making. In *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*, pages 287–291, 2012.

[9] Z. Xie, S. Huang, M. O. Ward, and E. A. Rundensteiner. Exploratory visualization of multivariate data with variable quality. In *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 183–190, 2006.