The Challenge of Missing and Uncertain Data (Poster)

Cyntrica Eaton and Catherine Plaisant
Human Computer Interaction Lab
University of Maryland, College Park
College Park, MD 20742
ceaton@cs.umd.edu, plaisant @cs.umd.edu

Terence Drizd
National Center for Health Statistics
3311 Toledo Road
Hyattsville, Maryland 20782
tad2@cdc.gov

1. Abstract

Although clear recognition of missing and uncertain data is essential for accurate data analysis, most visualization techniques do not adequately support these significant data set attributes. After reviewing the sources of missing and uncertain data we propose three categories of visualization techniques based on the impact that missing data has on the display. Finally, we propose a set of general techniques that can be used to handle missing and uncertain data.

2. Introduction

Information visualization presents an interesting paradox. While visual perception can be highly effective in the recognition of trends, patterns and outliers, the conclusions drawn as the result of such observations are only as accurate as the visualizations allow them to be. Therefore, to preserve the integrity of the data exploration process it is important to design visualization techniques that render data as accurately as possible and do not introduce misleading patterns. While this is an issue on a broader level, poor handling of missing values and data confidences is one specific aspect of data visualization that can have a negative influence on the quality of the data interpretation. Most tools available (especially research tools) cannot handle missing data and simply crash. The literature on visualization applications often reports on how the raw data has been preprocessed to "fill-in the blanks" or extrapolate data but users cannot see that the data was altered. Only rarely do tools attempt to make users aware of the presence of missing or uncertain information, e.g. [1,2,3]

We reviewed the sources of missing and low confidence data and propose a classification of visualization techniques based on the impact missing data has on the display and how likely users are to notice the existence of missing and uncertain data. Finally we propose a list of techniques that can be used to handle missing and uncertain data.

3. Sources of Missing and Uncertain Data

Because so many visualization tools work with data that can be represented in tabular form, we define a missing data point as an empty table cell. Generally, missing data is a result of the tools and procedures utilized during experimentation and constraints placed on the publication of data results, e.g. uncollected data, redefined data categories, data source confidentiality protection, and non-applicable multivariate combinations. Given the intrinsic collection and presentation influenced reasons behind missing data, avoiding missing values is nearly impossible, and the amount of missing data is likely to increase proportionally with the size of the set.

In most current visualization applications, however, missing data is either omitted from the display space, or presented in such a way that that it is indistinguishable from valid data. Consider the graph shown in Figure 1, as an example. Although the first three data points were actually missing, the preprocessing of the data filled the empty cells with zeros. Users are likely to interpret the diagram as showing the values to be low and stable then increasing sharply. This bias is likely to occur even if users are aware of what preprocessing took place.

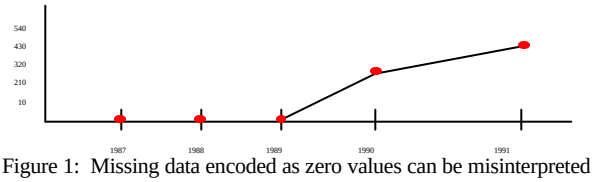

Figure 1: Missing data encoded as zero values can be misinterpreted

Confidence values are largely dependent upon the parameters of the experimentation process. Statistical sampling, sample size issues, flawed experimentation, and data estimation can all contribute to low confidence. While the missing data problem is more obvious in that a cell in a data set is actually empty, the confidence problem may even be more difficult to detect. The confidence interval may not be included at all in the data (it doubles the size of the dataset), or it may be difficult to present visually, and finally it may be difficult to comprehend for some users.

4. Classification of Visualization Techniques

We found three types of techniques in respect to the impact missing data has on the display. All visualizations use graphic objects to represent data items, and the position of those graphic objects on the display can be: 1) dedicated to the data item independently of the attribute values, 2) entirely a function of attribute values, or 3) a function of the item attribute values and the values of neighboring items.

An example of the first category ("dedicated") is a line graph in which the graphic object representing a data value is a dot with a dedicated X location. Other values in the data set have minimal influence on the graphic object. At most, the minimum and maximum values impact axis calibration. Chloropleth maps and techniques relying on ordering can fall in this category. For this type of visualization, if the data is missing and no object is displayed at the corresponding X position, the absence of data should be easily detected since users will be expecting to see a data point for each of the ordered values in the set (Figure 2).

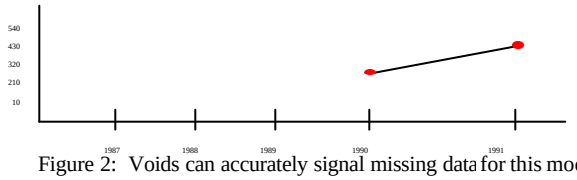

Figure 2: Voids can accurately signal missing data for this model

An example of the second category ("attribute dependent") is a scatter plot. In a scatter plot the position, color, and size of an object is based on the data item attribute values. If a data item is

missing, there is nothing in the display that clearly indicates a missing data value.

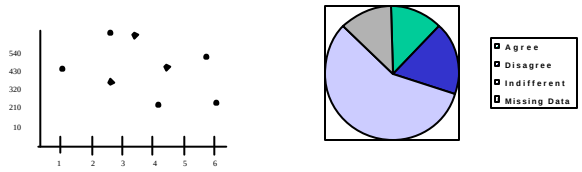

Figure 3: Voids can go undetected for these models

Examples of the third category ("neighbor dependent") are pie charts or Treemap. Here, the size and placement of the wedge or box representing the data item is a function of both the data item attribute values and neighboring items. If a data item is missing, simply omitting it from the display space will not only go unnoticed but it will also bias the appearance of other items. This is characteristic of all the space-filling techniques. In contrast the first two categories can be called neighbor-independent techniques.

Hybrid cases can be found. For example with parallel coordinates, a missing data item will go unnoticed (the position of the line is entirely dependent of attribute values) but a missing attribute value might be noticed as the location for that attribute is dedicated and the line can be drawn broken or connected to a separate location for missing values.

5. Possible Solutions and Directions

For both the neighbor-dependent and independent models, there are primarily three data visualization enhancements that could be used to provide effective indication of missing data and confidence intervals. They include 1) dedicated visual attributes, 2) annotation, and 3) animation. Dedicating visual attributes involves associating color, texture, shape, or any combination of these with data point appearance in order to indicate missing points or confidence ranges. Annotation, on the other hand, would allow users to gain further insight into missing and unreliable data through text or graphic information presented outside of the scope of data point appearance. Lastly, animation can provide a series of data display transitions that allow users to view several different perspectives in a short period of time. Animation can be helpful in adding and eliminating missing data clues based on the preference and/or intentions of the user. For example, users may be initially interested in observing the missing data points yet eventually hiding missing point indicators as set exploration goals change. Overall, the most effective way of using any of these enhancements is largely dependent upon the nature of the visualization paradigm.

For the visualizations in the first "dedicated" category, solutions abound as even a void can be noticed. Designers can use dedicated graphic attributes such as a special color or style to display an extrapolated value (e.g. a gray dot or a dotted line). They can also use annotation with a textual or graphic icon since there is dedicated space for it on the display, or they can use animation to first show only the data available then show the addition of the estimated data, possibly with a warning to users about the reason for the missing data. Similar techniques can be used to represent the uncertainty of the data. The color can become less intense with uncertainty, boxes or range bars can annotate the display, or animation can illustrate the possible variations of the display for min and max values. While both hatching and color ranges are both reasonably sound dedicated visual attributes that could be used to indicate associated confidence values, they could also be used to indicate the reasons why a given data point is unreliable. In either case, a particular hatching scheme or intensity would be mapped to a confidence value or a confidence influence and then incorporated into the display space to alert users accordingly. As stated before, these attributes should be carefully incorporated to ensure that the visualization does not become distorted, confounded, or ambiguous.

For the second category of visualizations ("attribute dependent"), designers have to rely on annotations to represent missing values. For example the number of missing items can be indicated on the side of the display, with possibly a list of names or partial representations when available. Hybrid cases exist where the data item may only be missing for some of the attribute values but not all of them. For example the X value can be known but the Y value missing, therefore specific annotations areas can be dedicated on the side that still represent the partial data. Ironically this category of visualization suffers from the opposite problem: data may sometime appear to be missing while in fact the graphic object is being hidden by another one. For uncertainty, dedicated graphic attributes, annotation or animation can be used. Data elements that vibrate in such a way that more stable data points indicate more confident measures might also provide users with the insights necessary to determine data point value dependability. Finally, methods like direct manipulation could provide the ability to filter data points upon demand based on user-defined confidence thresholds.

Neighbor-dependent visualizations are much more difficult to deal with as missing data is more likely to bias the interpretation of the rest of the data. Even choosing a default value for missing data has a significant impact on the display. Annotation is likely to be useful. Animation is intriguing. In a Treemap, as an example, the data can be shown without size coding with a dedicated color the extent of the missing data then animated to a size coded Treemap where the missing data is only indicated by a small fixed size of same color. The classic use of annotation for marking uncertainty (error bars) is a challenge for the neighbor dependent techniques. Animation of uncertainty is a challenge as elements interact with each others. Through direct manipulation, however, data analysts could be given the ability to filter the display space based on a user-provided range of confidence.

6. Conclusion

Dealing with missing data and uncertain data is a challenge for information visualization. We hope that our general classification of visualizations and techniques will help us to build effective prototypes that can be further tested to develop guidelines for designers.

7. Acknowledgement

This research was supported in part by the National Center for Health Statistics and NSF EIA 0129978

8. References

1. MacEachren, A. M., Brewer, C. A., and Pickle, L. 1998. Visualizing Georeferenced data: Representing reliability of health statistics. *Environment and Planning: A 30*, 1547-1561.

2. Twiddy, R., Cavallo, J., and Shiri, S. 1994. Restorer: A visualization technique for handling missing data. In *IEEE Visualization 94*, 212-216

3. Olston, C., and Mackinlay, J. 2002. Visualizing Data with Bounded Uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization*, 37-40