Visualizing Missing Data:
Classification and Empirical Study

Cyntrica Eaton[1], Catherine Plaisant[1], Terence Drizd[2]

[1]Human Computer Interaction Lab
University of Maryland, College Park, 20742
{ceaton, plaisant}@cs.umd.edu

[2]National Center for Health Statistics
Hyattsville, Maryland 20782

Abstract. Most visualization tools fail to provide support for missing data. We identify sources of missing and a categorization of visualization techniques based on the impact missing data have on the display: region dependent, attribute dependent, and neighbor dependent. We then report on a user study with 30 participants that compared three design variants. A between-subject graph interpretation study provides strong evidence for the need of indicating the presence of missing information, and some direction for addressing the problem.

1. Introduction

Information visualization provides an effective way for users to rapidly find trends in data and values of attributes of interest. The use of color, position, and shape contributes to helping users seeing patterns and outliers. Preserving the integrity of data exploration requires the use of visualization techniques that present data accurately without introducing misleading patterns or masking data properties. In particular, we believe that poor handling of missing and uncertain information can have a strong influence on users interpretation of the data (Fig. 1).
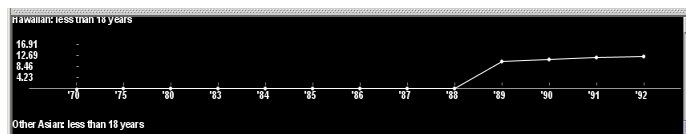
Fig. 1. In this figure the data seems to be stable, with a sharp increase starting in 88. Practically no data was collected until 89, so this interpretation is wrong.

When data is missing (e.g. there an empty cell in a data table), many tools will simply crash. Others will nicely inform users to "fix" the problem, which most users do by entering a value such as zero. As a result, it is often impossible for others to discriminate a value of zero from missing data. This paper categorize possible reasons

for data to be missing, differentiate three types of visualization techniques according to the impact missing data can have on the display and its interpretation, and report on a user study comparing three implementations.

2. Sources of missing data

As part of our research on making government statistics more accessible to the public (see Govstat project http://ils.unc.edu/govstat/) we found five main reasons for data to be missing:

Uncollected Data
The most trivial reason for missing data is that data was simply not collected. Equipment or sensors can malfunction, a survey can be misprinted, and files can be lost.

Data Source Confidentiality
Privacy protection can affect how findings are presented when publishing results of human-centric surveys or experiments. When the publication of a value might provide clues to the identity of individuals, that data must be omitted or presented aggregated at a higher level. For instance, when an organization publishes the average salaries of employees based on position and gender, the actual salary of the only female Vice President will be revealed. Publishing an empty cell is a solution, but if the number of male Vice Presidents is known, the aggregated data by position will also indirectly reveal her salary and should be omitted as well.

Redefined Data Categories
In statistical and demographic computation, data is often aggregated into classes or ranges [5]. Although aggregation is often necessary for efficient data presentation, problems arise when a class or range is redefined after data has been compiled. For example, U.S. population surveys did not allow people to select multiple races until the 2000 Census, so interracial population statistics are missing in years prior to 2000 even though citizens are counted in other categories. New definitions or discoveries of illnesses can also create complex missing data cases where studies of trends need to look at data across definition boundaries periods and understand the implications of the redefinitions.

Mutually Exclusive Multivariate Combinations
There are instances when combinations of data variables are impossible or highly improbable. Consider the example where the two variables of a dataset are age and cause of death by a firearm. Since it is not realistic to determine that a child of less than five years of age committed suicide, such category of data can be described as non-existing instead of having a value of zero.

Uncertainty deemed excessive
In some cases problems with small sample size, flawed methodology, and lack of data to use for estimation can contribute to high uncertainty for certain data values.   The authors of a study or report might decide to publish a simplified version of the dataset that does not include data with uncertainty over a certain threshold.

3. Classification of Visualizations

All visualizations use graphic elements to represent data, and we found that there are three categories of techniques (in respect to how much impact missing data has on the display) depending on how the position of the graphic elements is computed [21]. The position of the graphic elements 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 attributes 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. The values of other data items have no influence on the position of 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 then no object is displayed at the corresponding location, and the absence of data should be easily detected since users will be expecting to see a value there (Figure 2).
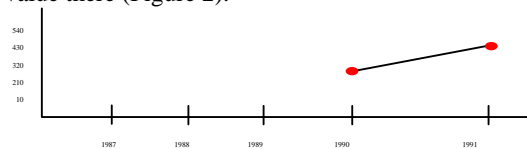

Fig. 2: Voids can be easily detected when there is a dedicated location for each data object

An example of the second category ("attribute dependent") is a scatter plot. In a scatter plot the position, color, and size of a graphical object is entirely based on the data item attribute values. If a data item is missing, there is nothing in the basic scatter plot display that indicates the existence of missing data value (Fig. 3).

Examples of the third category ("neighbor dependent") are pie charts and Treemap. Here, the size and placement of a 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 will not only go unnoticed but it will also bias the appearance of other items (Fig. 4). This is a characteristic of all the space-filling techniques.
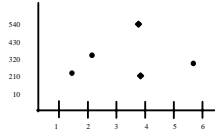
Fig.3   Attribute dependant example: In this scatter plot missing data is not noticeable
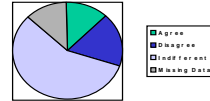

Fig.4   Neighbor dependant example: In this pie chart, not only is missing data not noticeable but it also biases the other data items (by making the other wedges larger than they should really be).

Cases of hybrid techniques exist as well. For example, with parallel coordinates, an omitted data item will go unnoticed because the position of the line is entirely a function of attribute values; but a missing attribute value might be noticed as the location for that attribute is dedicated and the line can be rendered broken or connected to a separate location for missing values.

We found three data visualization enhancements that could be used to provide effective indication of missing data and confidence intervals. They include:

- Dedicated visual attributes
- Annotation
- Animation

Dedicating visual attributes essentially involves associating color, texture, shape, or any combination of these with data point appearance in order to indicate missing values or specify 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 graphic element 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 temporary highlighting missing data, then adding estimated values, based on the preference and/or exploration goals of the user.

4. Related Work

Research in scientific visualization has given more attention than information visualization to missing data as well as uncertain data. In addition to specifically identifying sources of uncertainty, Pang et al. [12] discuss a classification of methods, present an overview of visual attributes that can be modified to indicate uncertainty. Pham and Brown [13] propose a list of relevant visual features that can be used to indicate data value imprecision (including hue, luminance, size, transparency, depth, texture, and blur) and present examples of "fuzzy" data. Cedlink and Rheingans [2], also providing clues and annotations such as grid lines. Restorer [16] use grayscale to indicate missing (and therefore estimated) data on color map. Djurcilov and Pang [4] discuss visualization techniques they used to analyze a sparsely populated meteorological dataset. Here a missing value is not an error but an indication that no phenomena were observable at a given point. They argue that missing data points

should not be estimated (as is usually the case), but presented in a way that alerts the user of "non-observation".   In contrast, Dybowski and Weller [5] address the problem of displaying missing information to the user by computing estimates and ranges.
MANET[17] and XGobi[15] attempt to make users aware of missing data and uncertainty.   They use complementary display that indicates the proportion of a missing data. For example in XGobi, a scatterplot is shown is two windows.   One contains the data, the other displays a shadow plot that indicates the data values that are complete, or missing the x, the y, or both attributes. Our exploration of the existing techniques highlights diversity of techniques and the challenge of providing visualization techniques that alert, yet do not distract. A common problem with the existing technique is that missing or uncertain data often ends up catching the eye the user more than the "good" data.

Empirical studies reporting on how users deal with missing or uncertain data are rare. Other studies involving graph interpretation (e.g. 18, 19, 20) assume a complete data set that did not include missing data. In the following section, we will discuss the pilot study we have conducted to better understand how users interpret simple graphs that include missing data.

5. Empirical study

Our goal was to study users' ability to compare data values and draw accurate conclusions about trends when data is missing, using three different displays. We wanted to be able to observe users dealing with missing data without making it obvious that missing data was the focus of our study, so each group of participants used only one of the three interfaces (i.e. we used a between subject design) and we asked them to answer some questions that involved missing data as well as some questions for which all the data was available.

Thirty people from the University of Maryland community participated in the study, 13 females and 17 males. Each participant was paid \$5.00 for taking part in the 20 minute study; and to improve motivation we also gave an extra \$5.00 to the participant with the highest accuracy and speed, in each of the three groups.

Microsoft Excel was used to create four separate time-sequence graphs. The graphs were then modified in a graphic presentation tool to transform them as necessary into one of the model variants. A tool was developed in C# to automate the presentation of the questions and displays, and collect time and preferences.

Figures 5-7 show three displays of the same data. In the *Misleading* display (Fig. 5), data values are encoded as 0. In the *Absent* display (Fig. 6) missing data is completely omitted from the display, and the line graph appear as broken when no data exist. The *Coded* display (Fig. 7), also omits missing data points but it add an icon on the next present data point in the series with that indicates why the prior data points are missing from the data set.
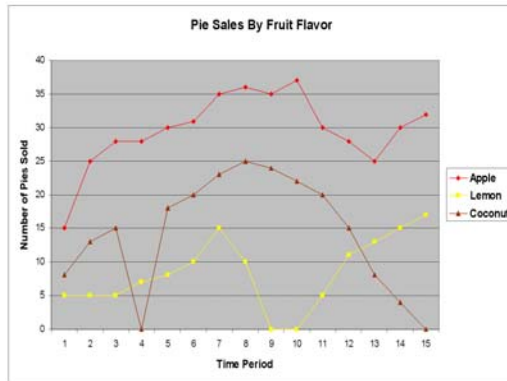
Fig. 5: *Misleading* Display - Missing data points are replaced by a default values (0).
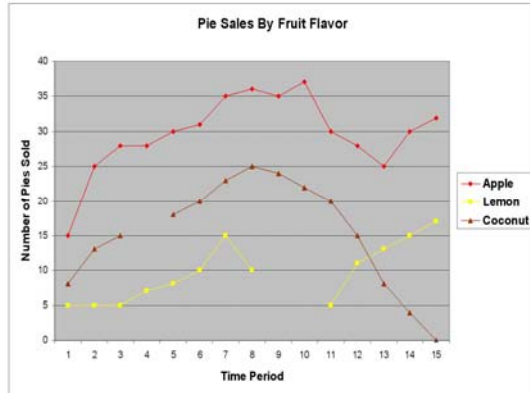

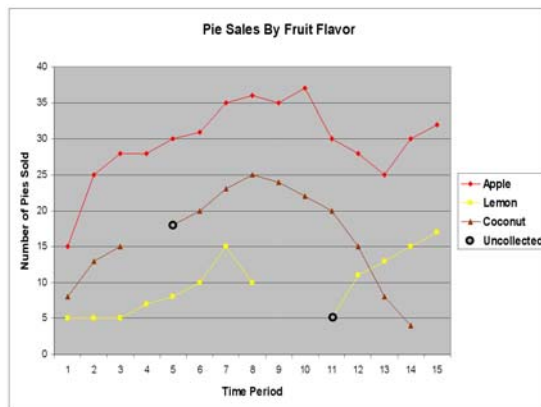Fig. 6 *Absent* Display - Missing data points are omitted.

Fig. 7 *Coded* Display - Missing data points are omitted, and the next valid point in the series has a mark which provides the reason why prior points are missing

We hypothesized that participants using *Coded* or *Absent* displays would be more accurate than participants using the *Misleading* display. We predicted that participants with *Absent* displays would have a shorter response time because they would have less information to digest, and that confidence and accuracy would be similar for users of both *Absent* and *Coded* displays and higher than *Misleading*. We thought that users would prefer the *Coded* version because it provides explanations.

Procedure

Participants signed the Informed Consent form and watched a brief slide show which explained a sample graph of the type they would be using during study. Instructions for answering comparison-based questions was provided. More specifically, to ensure uniformity in responses, participants were advised to answer questions of the form "Compare the value of *X* to *Y* at time t" in the form "*X* is greater/lower than *Y*". Next, each participant was given a brief overview of how the study would be executed.

They answered 13 questions. For each question the procedure was the same. The written question appeared on the screen. Once they had read the question and felt that they were ready to continue, they would click a button and a graph was displayed for five seconds, then hidden. The question reappeared along with a set of multiple-choice responses. For every question users could reply that they didn't have enough information to answer. After they had selected an answer (based on recall) and provided a confidence rating from 1 to 10, the graph reappeared and they were given a second opportunity to answer the same question while viewing the graph. The 1st answer measured the accuracy attained after a rapid glance at the graph , while for the final answer users had time to study the graph more carefully. After completing the study (using only one type of display: *Misleading, Absent* or *Coded*), users were shown examples of the other 2 displays and asked to choose the display they would prefer to use to answer the type of questions they had been given.

During the entire 20-minute session, the experimenter was seated beside the participant. She answered questions before the start of the experiment, observed participants and then asked clarifying questions after the experiment.

There were four types of questions: (with in parenthesis the notation used in the result charts)

- Value **Comparisons** where both points were **Present** (CP)
- **Trend**-related questions concerning only **Present** data (TP)
- Value **Comparisons** where one of the two points was **Missing** (CM)
- **Trend**-related questions involving **Missing** data (TM.)

The data was made-up but realistic, can carefully chosen so that it did not allow users to made conclusions based on their knowledge of the world, but based solely on the graph data they saw. For example data was about preferences of people from other planets, or imaginary illnesses. A complete list of sample graphs and questions used can be found at: www.cs.umd.edu/hcil/cyntricadata.html).

Results and discussion

Fig. 8 shows the average number of correct answers based on recall after a 5 second glance at the data. For each display there were 10 participants so a value of 10 means that all participants answered the question correctly every time, and a value of 0 means that none of the participants were able to answer the question correctly. For questions where all the data was present (CP and TP) users made a few mistakes, but the striking result is that none of the users were able to answer correctly to any of the questions involving missing data (CM and TM) using the *Misleading* display (remember that this is a commonly used way to present missing data). In each instance, participants indicated a definite trend or made a comparison between values as opposed to indicating that there was not enough information to answer the question. Even after being given more time to look at the display, they rarely changed their answers (Fig. 9). Users performed better with the *Absent* and *Coded* displays, but trends were still a problem, with great variability among users.

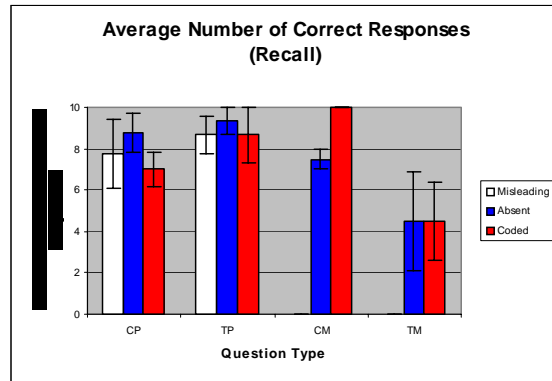**Average Number of Correct Responses (Recall)**

Fig. 8: The average number of correct responses based on recall after a 5 sec. glance at the data. The right 2 sets of bars show that users using the misleading display could not answer any of the questions correctly when missing data was involved (CM and TM).

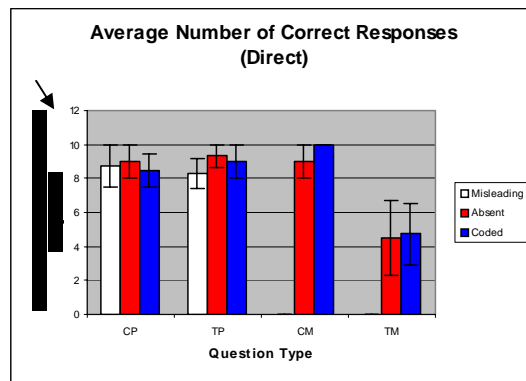**Average Number of Correct Responses (Direct)**

Fig. 9: The average number of correct final responses given while viewing the graph directly on the screen. Overall, users didn't change their answers when given more time.

Our hypothesis that participants with *Coded* and *Absent* displays would be more accurate than their counterparts using the *Misleading* displays was verified. The differences were significant when users compared between a missing value and a present data point ($p < 0.05$ for CM questions) and but less so when users have to describe a trend that incorporates missing values ($p < 0.10$ for TM questions). A closer look at the results showed that none of the participants using the *Absent* display answered two questions correctly. Both of these questions involved trend lines in which data was missing from the display. In both cases, the majority of users seemed to have constructed a confident opinion about the trend in the data based only on a few points of data shown in the display, as opposed to concluding that they did not have enough information to decide.

This supports our initial claim that poor indication of missing values can have a negative impact on data interpretation, but also suggests that even when missing data is indicated clearly users may not resist the temptation to find trends in partial data.

No significant differences between displays were found for confidence (Fig. 10 and 11). Users were confident in their answers. The average confidence value was nearly 8 for each of the models and for all of the questions, after 5 seconds and also when given more time.
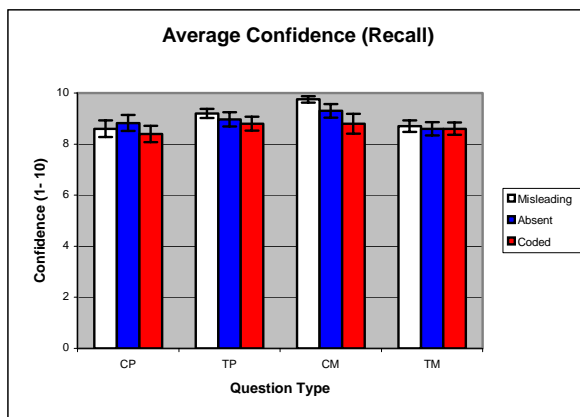

Fig. 10: Users where already very confident after viewing the graphs for 5 seconds, even when they made lots of errors (in CM and TM)

Concerning the time to answer no significant differences where found either, contradicting our hypothesis (Fig 12). For six of the thirteen questions answered, users with *Coded* displays had longer average response times. For four questions *Absent* displays had the longest response times while only two questions required more time to answer with the *Misleading* displays. Users of the *Misleading* displays seemed to behave as if the display was relatively straightforward and did not feel that they needed an extended period of time to ponder a response while some users of the other displays seem to hesitate more, but not all of them did so.

Eight users never changed their mind between the first answer and the final

answers, while seventeen users made one or two changes, and five users made three or more changes. On a category-by-category breakdown, of the eight participants who changed answers with the *Misleading* display, an average of two answers were modified with an average of one answer actually being changed to the correct answer. Users with *Absent* displays, changed an average of three questions, with an average of two modified to the correct response. Finally, users with *Coded* displays modified and average of two responses with an average of two actually being modified to the correct reply. Of the ten participants using the *Misleading* display, only one (a math major) commented at the end of the test that he was starting to suspect that missing data might have been an issue.
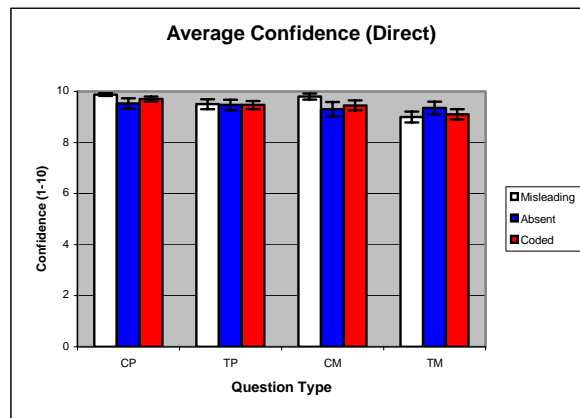

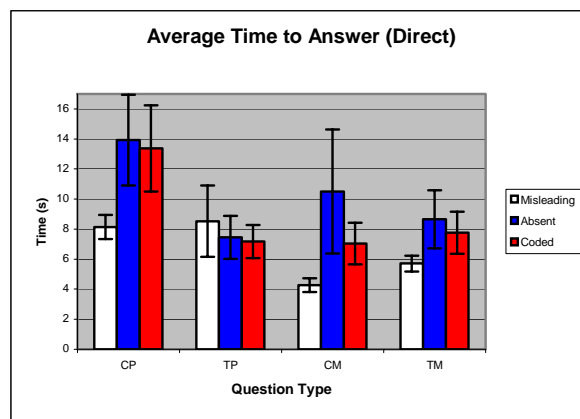Fig. 11: The final confidence level remains very high.

Fig. 12: The average time to give the final answer (while directly viewing the graphs). There was no significant differences.

When asked about their preference at the end of the test 27 users out of 30 selected the *Coded* display. They commented that they liked the idea of having more information available. Surprisingly two participants favored the *Absent* display over all three because they felt the *Coded* display was confusing. In the *Coded* display, the first present data point to appear after a series of missing points is encoded to convey the reason why previous data values were not available and this was found confusing. Finally, one user preferred the *Misleading* display because he liked to continuity of the graphs.

Conclusions

Accurately displaying missing and uncertain data presents an interesting challenge for information visualization. We hope that our general classification of visualization techniques will provide an useful basis for building and comparing techniques that represent missing data. Our study looked at how users interpret graphs with missing data. It suggests that users may not realize that data is missing when it is replaced by a default value. In real situations, the rate of error might be reduced because users can take advantage of their knowledge of the world to spot unlikely values. Furthermore, the study revealed that even if the missing data is noticeable, users are compelled to make general conclusions with partial data.

Participants preferred the *coded* display that provided additional information on the reason for the data to be missing. Some subjects voiced concern about the actual design of the *coded* display, suggesting that improvements could be made. Further studies of the impact of missing data on the more difficult cases of attribute dependant visualizations and neighbor dependant visualizations are needed as well.

Acknowledgments

References

1   Babad, Y.M., Hoffer, J.A. 1984. Even No Data Has a Value. *Communications of the ACM* 27(8), 748-756
2   Cedlink, A., Rheingans, P. 2000. Procedural Annotation of Uncertain Information. *IEEE Visualization*, 77-83
3   Chi, E. H. 2000. A Taxonomy of Visualization Techniques Using the Data State Reference Model. *Proceedings of Info Vis 2000*, 69-76
4   Djurcilove, S., Pang, A. 1999. Visualizing Gridded Datasets with large Numbers of Missing Values. *IEEE Visualization*, 405-408
5   Dybowski, R., Weller, P. 2001. Prediction Regions for the Visualization of Incomplete Datasets. *Computational Statistics16(1)*, 25-41
6   Ehlschlaeger, C. 1998. Exploring Temporal Effects in Animations Depicting Spatial Data Uncertainty. Available at: http://www.geography.hunter.cuny.edu/~chuck/aag98/
7   Gershon, N. 1999. Knowing What We Don't Know; How to Visualize an Imperfect World.

ACM SIGGRAPH Computer Graphics 33(3), 39-41

8   Healy, C. G., Booth, K.S., and Enns, J. T. 1996. High Speed Visual Estimation Using Preattentive Processing. *ACM Transactions on Human Computer Interaction 3(2)*, 107-135

9   Howard, D., MacEachren, A. 1996. Interface Design for Geographic Visualization: Tools for Representing Reliability. Available at: http://www.geovista.psu.edu/publications/others/howard/howmac96.html

10   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.

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

12   Pang, A. T., Wittenbrink, C. M., Lodha, S.K. 1996. Approaches to Uncertainty Visualization. Technical Report UCSC-CRL-96-21, University of California, Santa Cruz.

13   Pham, B., Brown, R. 2003. Visualization: An Analysis of Visualization Requirements for Fuzzy Systems. *First International Conference on Computer Graphics and Interactive Techniques*, 181-187

14   Shniederman, B. 1996. The Eyes Have It: A Task by Data Type Taxonomy of Information Visualizations. *IEEE Visual Languages*, 336-343

15   Swayne, D. F., Buja, A. 1998. Missing Data in Interactive High-Dimensional Data Visualization. *Computational Statistics 13(1)*, 15-26

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

17   Unwin, A. Hawkins, G., Hofmann, Siegl, B. 1996. Interactive Graphics for Data Sets with Missing Values – MANET. *Journal of Computational and Graphical Statistics 5(2)*, 113-122

18   Beichner, R. 1994. Testing Student Interpretation of Kinematics Graphs. *American Journal of Physics 62*, 75-762

19   Roth, W., Gervase, M.B. 2003. When Are Graphs Worth Ten Thousand Words? An Expert-Expert Study. *Cognition and Instruction 21(4)*, 429-473

20   Brassuer, L. 1999. The Role of Experience and Culture in Computer Graphing and Graph Interpretive Processes. *Proceedings of the 17th annual international conference on Computer documentation*. 9-15

21   Eaton, C., Plaisant, C., Drizd, T., The Challenge of Missing and Uncertain Data Poster in the *Visualization 2003 Conference compendium*, IEEE, (2003) 40-41