Integrating Data and Quality Space Interactions in Exploratory Visualizations [*]

Zaixian Xie, Matthew O. Ward, Elke A. Rundensteiner and Shiping Huang
Computer Science Department
Worcester Polytechnic Institute
Worcester, MA, USA
{xiezx,matt,rundenst,shiping}@cs.wpi.edu

Abstract

Data quality is an important topic for many fields because real-world data is rarely perfect. Analysis conducted on data of variable quality can lead to inaccurate or incorrect results. To avoid this problem, researchers have introduced visual elements and attributes into traditional visualization displays to represent data quality information in conjunction with the original data. However, little work thus far has focused on creating an interactive interface to enable users to explicitly explore that data quality information. In this paper, we propose a framework for the linkage between data space and quality space for multivariate visualizations. Moreover, we introduce two novel techniques, quality brushing and quality-series animation, to help users with the exploration of this linkage. A visualization technique specifically designed for the quality space, called the quality map, is proposed as a means to help users create and manipulate quality brushes. We present some interesting case studies to show the effectiveness of our approaches.

1. Introduction

Data quality information, such as accuracy, reliability, and uncertainty, are important properties of real data. To separate two kinds of information, the original data and its quality attributes, we define two spaces, namely *data space* and *quality space*. The former consists of the original data, and the latter consists of data quality information.

In order to draw correct conclusions even if some of the data is of low data quality, numerous research efforts have focused on visualizing data quality along with the original data. However, most of these efforts did not explicitly support quality-related interactions. As we know, the goal of visualization is not merely to translate data into various display forms, but it is equally and possibly more important to provide an interactive interface to enable users to retrieve the information from the display, find patterns in data, and draw conclusions. Therefore, explicit quality-based interactions would make the exploration tools more quality-aware, and thus more effective. Quality-aware interactions would let users sift through quality space itself or explore the linkage between data space and quality space.

In our prior research [27], similar to other research efforts, we have integrated quality measures into traditional multivariate visualizations. However, as indicated above, we focused on visualizing data quality, but did not make the interactions quality-aware. Thus our current goal is to create an interactive system to help analysts retrieve and utilize quality information in multivariate datasets in an intuitive and efficient manner.

In quality-extended datasets, there are numerous tasks related to interactively selecting subsets. For example, users might want to highlight datapoints with high quality to draw confident conclusions, or focus on those with low quality to identify the potential cause of low quality. Brushing is a widely used mechanism for the fundamental task of interactively selecting a subset of data in dynamic data visualization applications [4]. To extend traditional brushing to become more quality-aware, we need to answer two questions: (1) how to visualize the quality space; (2) how to help users select subsets in data space and quality space and demonstrate the linkage between the two spaces. Starting from these two questions, our primary contributions in this paper include:

- *Visualization of quality space*: To visualize the quality space, we propose a customized visualization technique, the *Quality Map*.
- *Interactions between data space and quality space*: We define and implement linkage operations between visualizations in data space and quality space. The interaction system allows users to define or manipulate

[*] This work is supported under NSF grants IIS-0119276, IIS-00414380 and a grant from the NSA.

a brush in either space and observe the corresponding selected datapoints in the other space.

- *Multiple brushes and logical combinations*: We allow users to define more than one brush in quality space. The resulting subsets in data space can be combined using logical operations.

- *Procedural brushing by quality-series animation*: We associate one of the quality measure dimensions with the time attribute and create a type of time-series animation, which we term quality-series animation. Since each frame shows a data subset in terms of quality ranges that are automatically changing over time, we call this mechanism *procedural brushing*.

The remainder of this paper is organized as follows: In Section 2, we review existing techniques for quality-related visualization and brushing. In Section 3, we introduce the quality space and present an overview of quality-aware interactions. Section 4 describes the *quality map*, a technique specially designed for the visualization of the quality space. Section 5 presents the linkage between data space and quality space. Sections 6 and 7 describe the details for implementing the linkage between the two spaces using two techniques: N-dimensional brushes and quality-series animation. In Section 8, we investigate an interesting dataset and show the effectiveness of the proposed quality-aware interactions. Section 9 gives a summary and possible future research directions.

2. Related work

In this section, we discuss recent efforts of researchers in data quality visualization and interactive brushing, the two technologies we aim to combine.

Data Quality Visualization: We can find numerous research efforts regarding data quality visualization in the recent literature. Important topics related to our work include the definition and modeling of uncertainty, missing data visualization, and uncertainty visualization.

A report by NIST [20] divided uncertainty into two categories, named Type A and Type B. The former defines the uncertain value by a peak value and a potential distribution, while the latter gives a precise lower and upper bounds to convey the uncertainty. Olston and MacKinlay [15] called these two types statistical uncertainty and bounded uncertainty. However, we have found it to be more convenient to use a scalar value to present the certainty degree [26, 3], although some more complex representations are possible.

XGOBI [17] and MANET (Missing Are Now Equally Treated) [22, 9] are data visualization tools designed to handle missing data. They replace missing fields with estimated values, to which indicators (e.g. different colors or positions) are attached showing that these values are substitutions. The GIS community has produced a large amount of research regarding data quality issues, focusing on uncertainty definition, modeling, computation and visualization [11, 12]. They discussed a lot of possible graphical variable mappings to represent uncertainty, including color, hue, texture, fog, animation, and flashing. Wittenbrink, Pang, and Lodha [26, 16] proposed techniques for visualizing uncertainty found in vector fields. Many mappings of uncertainty degree to glyph attributes were developed and evaluated, including adding glyphs, adding geometry, modifying geometry, modifying attributes, animation, sonification, and psycho-visual approaches.

However, these authors did not explicitly discuss how to create an interactive interface to allow users to explore the quality-extended dataset. Our goal is to fill this gap.

Brushing: Discussions about brushing exist in some very early systems, such as PRIM-9 [21] by Tukey, Fisherkeller and Friedman. This system allowed users to interactively select a region of the data display, although they did not call it *brushing*. The principles of brushing were introduced and discussed extensively by Becker and Cleveland [2]. They presented several different ways to select a data subset in scatterplot matrices. Many concepts and operations, such as definition, creating, changing and moving of the brush, can be applied to other multivariate visualizations.

In more recent work, researchers discussed the taxonomy, semantics and operations for brushing. Wills presented a comprehensive set of possible selection operations, and then identified five most useful selection operations (replace, intersect, add, subtract, toggle) and possible combinations [25]. Chen proposed a conceptual model, called compound brushing, to model brushing techniques [4]. Brushing techniques were modeled as higraphs with five types of basic entities: data, selection, device, renderer, and transformation. A flexible visual programming tool was built to make use of these entities and create combinations of brushes.

Brushing has been incorporated into many visualization tools. Velleman introduce brushing to many plots other than scatterplot matrics, such as histograms, bar charts, and so on [23]. Martin and Ward implemented N-dimensional brushing in the multivariate visualization package XmdvTool [13]. Up to four brushes are supported by XmdvTool. Users can do union, intersection and complement operations on selected subsets. Moreover, XmdvTool supports the notion of fuzzy brushing, which consists of selection via ramped brushes. Fua et al. introduced structure-based brushing to perform selection in hierarchically structured datasets [7, 8]. Swayne et al. proposed and implemented linked brushing in GGobi [18, 19]. In this open source visualization program for exploring high-dimensional data, users can link multiple views for one dataset such that brushing points in one view can cause the same points to change colors in other views. Our focus is to expand brushing into the quality space.

3. Overview of Quality Space and Interactions

In this section, we first define the structure of quality space, and then describe the concept of quality-aware interactions.

3.1. Quality Space

To describe different kinds of quality issues in multivariate data, we introduce three types of quality measures: data value quality, record quality and dimension quality. Each entry of these quality measures represents the uncertainty degree of each data value, each record and each column, respectively. Without loss of generality, we employ a scalar value to represent each entry and normalize it to the range from zero (lowest quality) to one (perfect quality).

More formally, we define the quality measures for a multivariate dataset D as: [27]

$$(Q_V, Q_R, Q_D) \tag{1}$$

which forms the quality space. Note that Q_V, Q_R, Q_D denote the data value quality, record quality and dimension quality. If the number of dimensions and records in dataset D is n and m, then

$$\begin{aligned} Q_V &= \left\{ \begin{array}{llll} v_{11} & v_{12} & \dots & v_{1n} \\ v_{21} & v_{22} & \dots & v_{2n} \\ \dots\dots & & & \\ v_{m1} & v_{m2} & \dots & v_{mn} \end{array} \right\} && (2) \\ Q_R &= (r_1\ r_2\ \dots\ r_m)^T && (3) \\ Q_V &= (d_1\ d_2\ \dots\ d_n) && (4) \end{aligned}$$

where $0 \leq v_{ij}, r_j, d_i \leq 1$.

Note that our definition is only structural and not semantic. Analysts in a specific application domain must determine what quality means in their domain and then map their measures appropriately into the entries in Q_V, Q_R and/or Q_D. Moreover, a particular domain may need only one or two of these three kinds of data quality.

3.2. An Overview of Quality-Aware Interactions

Figure 1 shows our approach for interactions on quality-extended datasets. First, data space and quality space are visualized. For the data space, we employ traditional visualization techniques. For the quality space, we designed *Quality Maps*, a visualization technique specially designed to represent data quality information. Then we introduce two kinds of interactions:

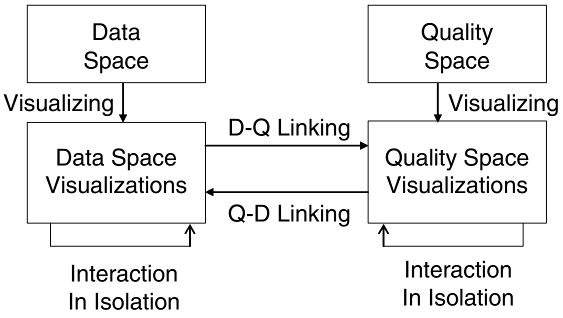


Figure 1. Interactions within and between data space and quality space

- *Linkage Interactions*: Q-D linking allows users to select a subset in quality space and then observe the corresponding datapoints in data space. D-Q linking is the inverse process. Details are discussed in Section 5.

- *Interaction in isolation*: Either data space or quality space visualizations can be analyzed independently from each other. Various brushing techniques have been introduced for multivariate visualizations [4, 13, 24]. We can apply such techniques to these two spaces in isolation.

4. Quality Maps

In this section, we propose one new technique for quality space visualization, which we call *Quality Maps*. One design alternative might be to treat the combination, $\{Q_V, Q_R\}$, corresponding to the data value quality and record quality from Equation (1), as a single tabular dataset, so any multivariate visualization technique can be applied to quality space. However, for quality-related tasks, we often need to select a specific quality range. For example, analysts might select quality measures from 0.7 to 1.0 to observe the distribution of these measures, or find patterns in the corresponding datapoints. Traditional multivariate visualizations, such as scatterplot matrices and parallel coordinates, do not, in our experience, provide an effective solution. Thus it was necessary to consider alternative designs.

To help explain our approach, we give an example using a variation of the *iris* [6] dataset modified by adding noise. The data value quality was computed based on the difference between the original value and the modified one.

In Figures 2 and 3, we present two types of *Quality Maps*, namely *Stripe Quality Maps* and *Histogram Quality Maps*. In Figure 2, each stripe in the *data value quality*, *record quality* and *dimension quality* sections corresponds to one quality value. The brightness of each stripe reflects

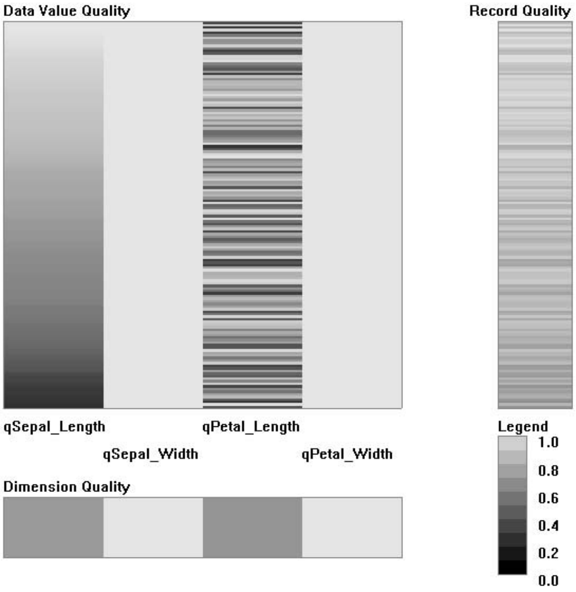


Figure 2. ***Stripe Quality Map*** **for quality space. Datapoints are sorted by the value quality of the first dimension. Dimensions 2 and 4 have perfect quality.**

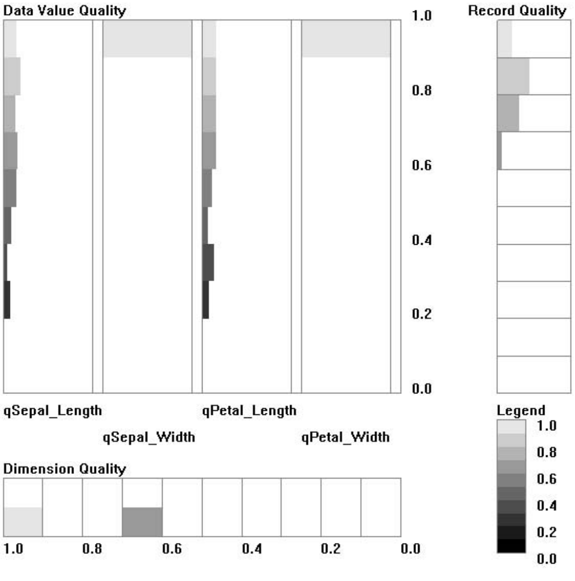


Figure 3. ***Histogram Quality Map*** **for quality space.**

the quality measures as shown in the legend at the bottom right corner. To help users explore the distribution of quality measures, we provide two kinds of sorting: (1) Sorting all of datapoints in the order of value quality of one dimension or record quality, or (2) sorting all columns in the order of dimension quality values. In Figure 3, we use $n + 2$ histograms corresponding to each value quality dimension, record quality and dimension quality. Here n is the number of dimensions in the original dataset.

Discussion of Quality Displays After some testing with several sample datasets, some advantages and disadvantages about these two techniques became clear. These are as follows:

- *Histogram Quality Maps* are more suitable for tasks that involve retrieving distributions of quality measures as compared to *Stripe Quality Maps*.
- In *Stripe Quality Maps*, users readily perceive the relationship among quality dimensions. For example, based on Figure 2, we can conclude that the quality of the third dimension is not correlated with that of the first dimension. In other words, they are independent.
- In *Stripe Quality Maps*, users can retrieve the distribution of the sorted dimensions more easily than that of unsorted dimensions. For example, in Figure 2 it is much easier to find the percentage of quality values in the [0.8,1] on the first dimension than in the third dimension, since the data is sorted by the first dimension.

5. Linkage between Data Space and Quality Space

In this section, we propose a new operation, *Value-Attribute Linking*, to link data space and quality space. We can regard it as a new kind of brushing. It has two directions.

- *Linking from quality space to data space*: As shown in Figure 4 (a), when users select a range in quality space, all datapoints falling into this quality range will be highlighted in the data space. Note that these datapoints are not necessarily contiguous in the display. We call this linking *quality brushing*. Section 6 presents implementation details.
- *Linking from data space to quality space*: Figure 4 (b) shows the link in another direction. When users highlight a subset in data space using traditional N-dimensional brushing, the quality measures of datapoints in this subset are highlighted in the quality space.

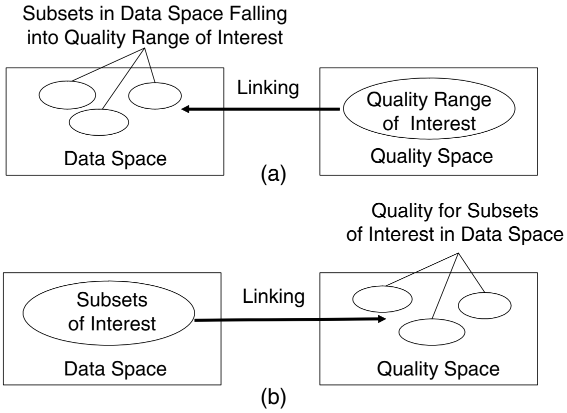


Figure 4. Bidirectional linking between data space and quality space.

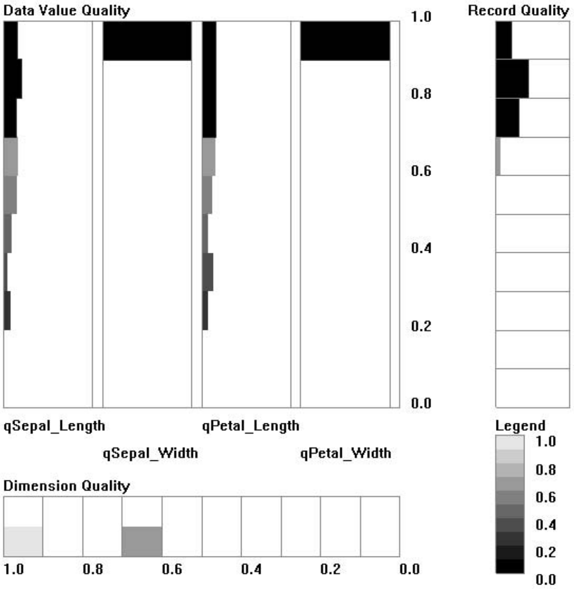


Figure 5. Brushing on the *Histogram Quality Map* (black regions). The resulting brushing in data space is shown in Figure 6

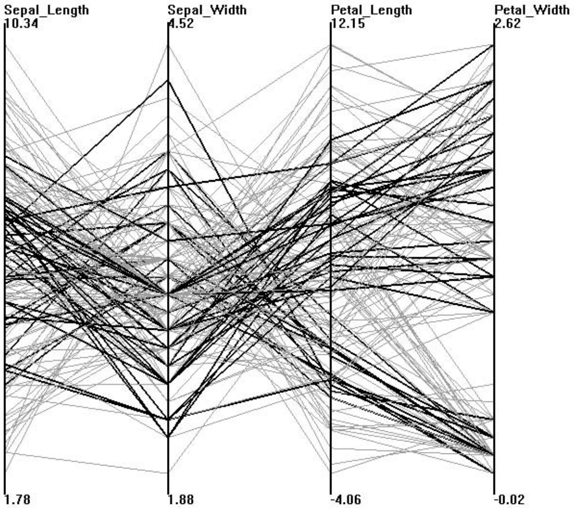


Figure 6. The resulting brushing in data space linked via the quality space brushing shown in Figure 5.

Figures 5 and 6 show the first type of linking. In Figure 5, users select high values (high confidence) in the quality space (the black regions). The linked datapoints in the data space are highlighted in black as shown in Figure 6. Because these datapoints have high quality, we can focus on the darker lines in Figure 6 to draw reliable conclusions. For instance, we can identify the roughly positive correlation between the dimensions *Petal_Length* and *Petal_Width* since the darker lines between these two dimensions are approximately parallel.

From Figure 6, we can see that datapoints with low quality (light-gray lines) normally have higher values on dimensions *Sepal_Length* and *Petal_Length*. To confirm this finding, we use the second linking from data space to quality space. We show the brushing on data space in Figure 7 to select higher values on dimension *Petal_Length*. Then we use the linked brushing to derive Figure 8 and show the distribution of quality measures for datapoints highlighted in Figure 7. Let's compare Figures 8 and 3. As we know, the latter describes the same distribution as Figure 8 but for the whole dataset. We can see that a greater percentage of quality measures in Figure 8 fall into low quality ranges than Figure 3. So an obvious conclusion is that datapoints highlighted in Figure 7 have lower value quality measures than other datapoints. Our finding is thus confirmed.

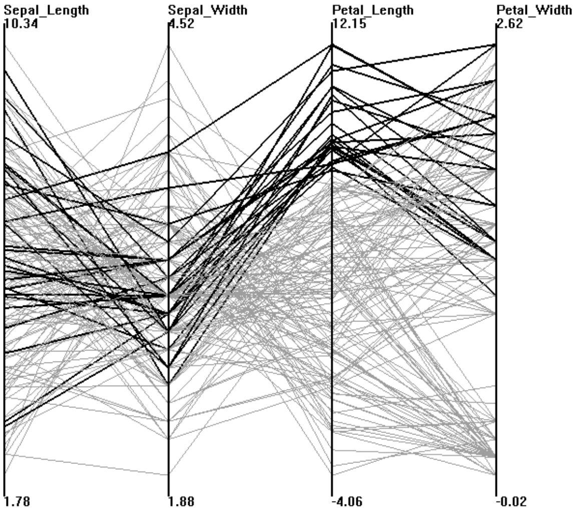


Figure 7. Datapoints with high values on dimension *Petal_Length* are selected. The linked quality space is shown in Figure 8.

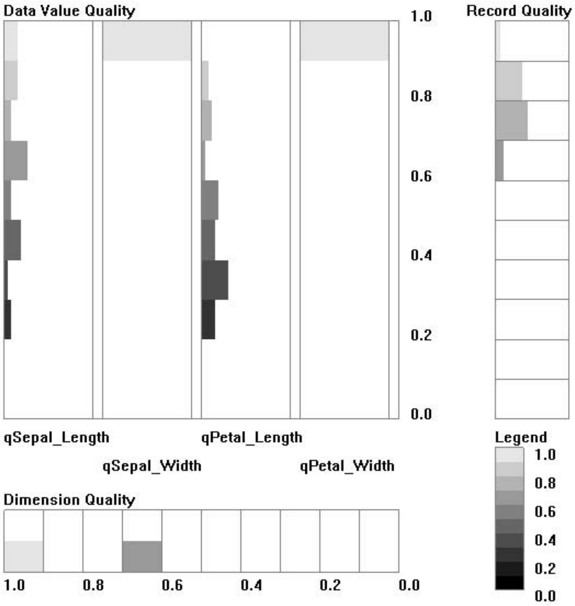


Figure 8. The distribution of quality measures for datapoints highlighted in Figure 7.

6. Quality Brushing with N-dimensional Brushes

In this section, we discuss the details of quality brushing. and the linkage from the quality space to the data space. We will first define quality brushing as a variation on Martin and Ward's N-dimensional brushes [13], and then explain our implementation.

6.1. Definition of n-dimensional quality brushing

Our quality brushing is somewhat different from Martin and Ward's N-dimensional brushes. In their definition, whether a datapoint is highlighted or not depends on its N-dimensional values. An N-dimensional brush is a hyperbox that can be defined by an n-tuple

$$([S_1, E_1], [S_2, E_2], ..., [S_n, E_n]) \tag{5}$$

where $[S_j, E_j]$ denotes the start and end values on dimension j, which specify a range in which users are interested. Such a brush results in a selected subset of the original dataset:

$$\{(a_{i1}, a_{i2}, ..., a_{in}) | S_j \leq a_{ij} \leq E_j, 1 \leq j \leq n\} \tag{6}$$

where the tuple $(a_{i1}, a_{i2}, ..., a_{in})$ denotes a datapoint.

However, the coverage of a quality brush depends on attributes of the datapoint, namely, the quality measures. Similar to the above definition, we can describe our quality brushing as a quality hyperbox, an $(n + 2)$-tuple:

$$([Sv_1, Ev_1], [Sv_2, Ev_2], ..., [Sv_n, Ev_n], \\ [Sr, Er], [Sd, Ed]) \tag{7}$$

where $[Sv_j, Ev_j]$ is a subrange of data value qualities on dimension j, and $[Sr, Er]$, $[Sd, Ed]$ are ranges of the record quality and dimension quality respectively. Based on the brush in (7), our quality brushing would highlight the datapoints in the following subset

$$\pi_{j_1, j_2, ..., j_k} S \tag{8}$$

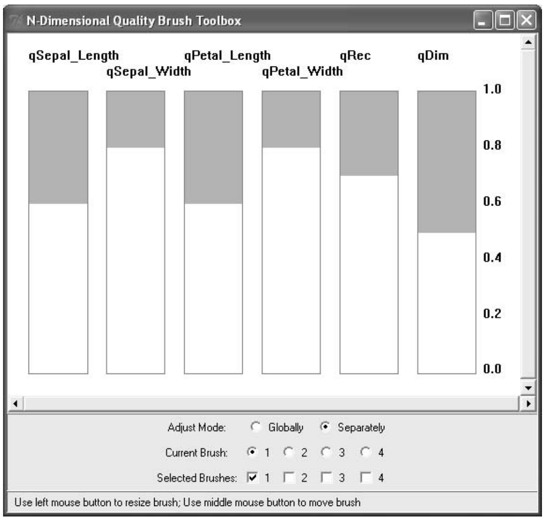
where $\{j_1, j_2, ..., j_k\}$ is the subset of $(1, 2, ..., n)$, satisfying $[Sd \leq d_{j_1}, d_{j_2}, ..., d_{j_k} \leq Ed]$. In other words, $j_1, j_2, ..., j_k$ are those dimensions whose dimension quality measures fall into the quality range specified by the user. The operator π selects only those dimensions as the output. S can be represented by:

$$S = \{(a_{i1}, a_{i2}, ..., a_{ij}, ..., a_{in}) | Sv_j \leq v_{ij} \leq Ev_j, \\ Sr \leq r_i \leq Er, 1 \leq j \leq n\} \tag{9}$$

where v_{ij} and r_i are scalar measures for data value quality and record quality respectively. The subset S contains only those datapoints with appropriate quality measures of interest in terms of a user's selection.

6.2. Implementation of n-dimensional quality brushing

Figure 9 shows our toolbox being used to define a quality hyperbox. If the number of dimensions in the original dataset is n, there are $n + 2$ rectangles in the toolbox. The first n sliders correspond to the data value quality for each dimension, and the last two denote the record quality and the dimension quality. The quality hyperbox in Figure 9 is $([0.6, 1], [0.8, 1], [0.6, 1], [0.8, 1], [0.7, 1], [0.5, 1])$. Through this quality brush, users can select and highlight the datapoints with high certainty. In this toolbox, we allow users to click on the rectangle to change the range for each quality dimension. To extend the usefulness of this toolbox, we integrate two features, *global adjustment* and *multiple quality brushes*.



Adjust Mode: ○ Globally ● Separately

Current Brush: ● 1 ○ 2 ○ 3 ○ 4

Selected Brushes: ☑ 1 □ 2 □ 3 □ 4

Use left mouse button to resize brush; Use middle mouse button to move brush

Figure 9. quality brushing definition toolbox

Global adjustment: This feature means that users can adjust the quality range globally. That is to say, if users set the adjust mode to *globally*, they only need to adjust one dimension, and all other dimensions will keep the same range as the one being adjusted.

Multiple quality brushes: We allow users to define up to four quality brushes simultaneously and decide which brushes are enabled. Assume that four brushes are B_1, B_2, B_3 and B_4, and the corresponding highlighted subsets are S_1, S_2, S_3 and S_4. The datapoints in these four subsets are rendered with different colors. The merging of brushes via a union operation is available. Although more operations, such as intersection and complement, are possible, we did not implement them yet since we felt that the union operation is the most useful for most quality-related explorations. More operations will make the system more complex. For example, B_1, B_2, B_3 and B_4 are set to the quality ranges $(0.9, 1.0)$, $(0.8, 0.9)$, $(0.7, 0.8)$, $(0.6, 0.7)$. Using our system, users can get S_1, $S_1 \cup S_2$, $S_1 \cup S_2 \cup S_3$, and $S_1 \cup S_2 \cup S_3 \cup S_4$ step by step. In this process, users can increase the quality range gradually and thus select more and more datapoints.

7. Procedural Brushing by Quality-series Animation

In this section, we make use of animation to integrate quality information into traditional multivariate visualization to facilitate quality-related exploration. As we know, an animation consists of a series of frames $(F_1, F_2, ..., F_n)$. F_i is a function of time T. Images change over time to reflect movement of objects. However, not all multivariate datasets have time attributes; moreover, our main goal is to reflect the relationship between data values and quality measures. Therefore, our basic idea is to regard the quality levels as time T. In our animation, we make quality measures change gradually from one value to another value with a small interval, for instance, in the series 1.0, 0.99, ..., 0.0. For each quality measure in this series, namely, q_{now}, we create a subset of the original dataset and render this subset as a frame. For example, we can include all datapoints with the record quality greater than q_{now}. The time interval between two adjacent frames might be bigger than typical animations to leave time for users to examine each frame carefully. To conceptualize our method, we introduce several constants, q_0, q_1, Δ_q, and a function F_{sq}. Three constants are used to generate the quality series, q_0,$q_0 + \Delta_q$,$q_0 + 2\Delta_q$,...,q_1. F_{sq} can generate a subset in terms of the current quality measure, q_{now}. Note that we do not apply the dimension quality in this function, since we do not want to have dimensions flashing from one frame to the other. Such a case will likely be too confusing because most multivariate visualizations will have a significant change if the number of dimensions changes. In the following text, we can see that different definitions for the function F_{sq} can provide us different views of the quality-extended dataset.

Here we propose two definitions for the function $F_{sq}(q)$, namely *slot view* and *aggregation view*.

slot view: We define the function $F_{sq}(q)$ as

$$\begin{aligned} F_{sq} : q \rightarrow \{(a_{i1}, a_{i2}, ..., a_{in})| \\ q \leq F_q(v_{i1}, v_{i2}, ..., v_{in}, r_i) \leq q + |\Delta_q|\} \end{aligned} \tag{10}$$

where F_q returns a quality measure for the datapoint $(a_{i1}, a_{i2}, ..., a_{in})$. It can be the data value quality on an arbitrary dimension v_{ij} $(1 \leq j \leq n)$, record quality measure r_i or even a statistical resulting from these measures, such as the average, minimum, maximum values and so on. Using the definition in (10), each frame is like a quality slot, only showing the datapoints with F_q falling into the range

$[q, q + |\Delta q|]$. For our sample dataset, if we let $q_0 = 0.9$, $q_1 = 0.2$, $\Delta_q = -0.1$, and $F_q(...) = v_{i1}$, we can get an animation having 8 frames. Each frame is shown in Figure 10.

aggregation view: If we define $F_{sq}(q)$ as

$$F_{sq} : q \rightarrow \{(a_{i1}, a_{i2}, ..., a_{in})| \\ q \leq F_q(v_{i1}, v_{i2}, ..., v_{in}, r_i) \leq q_0\} \ (\Delta_q < 0) \tag{11}$$

or

$$F_{sq} : q \rightarrow \{(a_{i1}, a_{i2}, ..., a_{in})| \\ q_0 \leq F_q(v_{i1}, v_{i2}, ..., v_{in}, r_i) \leq q\} \ (\Delta_q > 0) \tag{12}$$

we can get an animation to show more and more datapoints in contiguous frames.

8. Case Studies

To produce a dataset with quality attributes, we used a dataset, *Horse-Colic* from [14]. 30% of the values are missing. The original dataset has 28 columns. We selected 7 of them to illustrate our technique more clearly. This dataset describes horses having suffered colic using some pathologic parameters. Starting from this dataset, the quality measures can be derived by imputation algorithms, where the missing values are replaced with synthetic values. Below is the detailed description.

We employed a multiple imputation algorithm [1, 10] to generate estimated values for missing values. This algorithm repeats the imputation process more than once, producing multiple complete data sets until the estimates converge. Therefore, we can get p values for each missing value, namely $y_k(1 \leq k \leq p)$, if we repeat the imputation process n times. The formula to calculate data value quality for this missing value is given by

$$v = 1.0 - \frac{\delta_y}{X_{max} - X_{min}} \tag{13}$$

Here X_{max} and X_{min} are the maximum and minimum values of the dimension on which the missing value is, and δ_y is the standard deviation of $y_k(1 \leq k \leq p)$. If the initial estimated value is close to the final imputed value, δ_y is small and V is close to 1.0. It shows that the imputed value is reliable. Otherwise, the imputed value is not reliable. If $\delta_y > X_{max} - X_{min}$, we set the quality measure to 0.0.

We compute record quality and dimension quality measures by the following formulas:

$$r_i = min\{v_{ij}, 1 \leq j \leq n\} \tag{14}$$

$$d_j = \frac{\sum_{i=1}^{m} v_{ij}}{m} \tag{15}$$

Here n is the number of dimensions, and m is the number of records.

Figure 11 shows the dataset *Horse-Colic* in Parallel Coordinates. We can observe the distributions of quality measures in Figure 12. The columns *Respiratory* and *PH* have many values of low quality resulting from the imputation algorithm, which can make the patterns in Figure 11 not so obvious. If we use a quality brush to highlight datapoints with high quality, we generate Figure 13. If we focus on datapoints of high quality (darker lines), we can find some interesting patterns. For example, a lower degree of pain (*Pain*) normally corresponds to a lower respiratory rate (*Respiratory*) and a smaller number of red cells (*Redcell*).

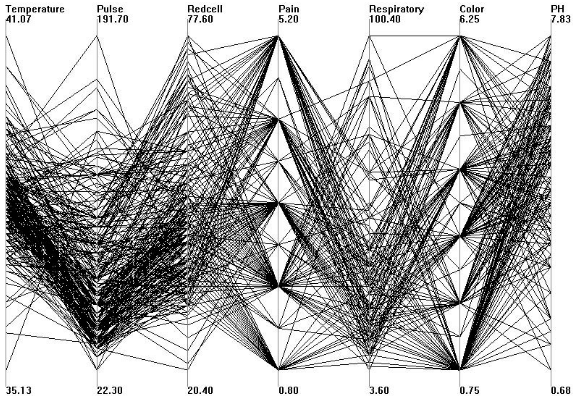


Figure 11. The dataset *Horse-Colic* in Parallel Coordinates.

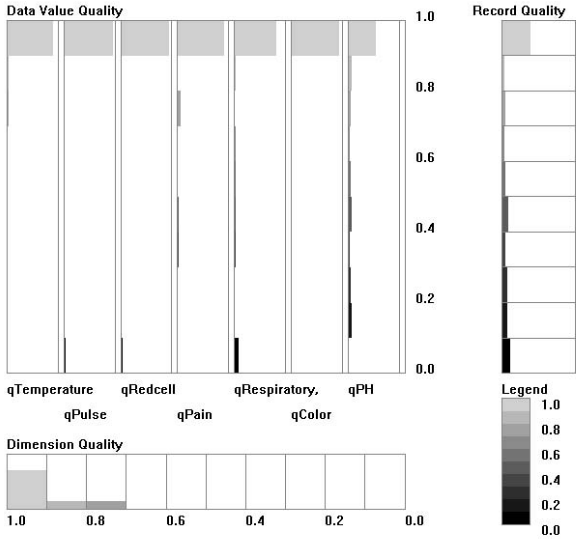


Figure 12. *Histogram Quality Map* for the quality space for the dataset *Horse-Colic* .

In Figure 13, we find that most of the lines crossing higher values on column *Respiratory* are light-gray lines,

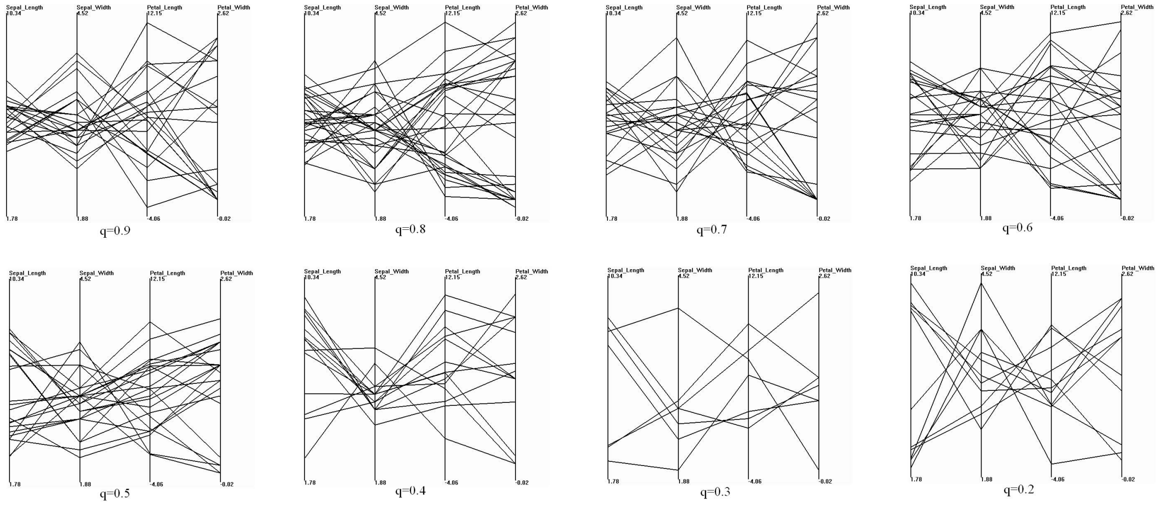


Figure 10. Quality-series animation.

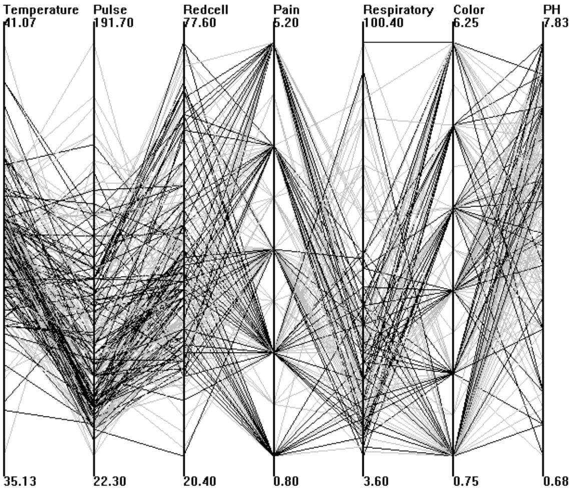


Figure 13. The resulting parallel coordinates of the dataset *Horse-Colic* by selecting higher quality measures in the quality space. Darker lines correspond to the datapoints of high quality.

which indicate they have low quality measures. To confirm this possible relationship between data space and quality space, we use the linking from data space to quality space. In Figure 14, we create a data quality brush to select higher values on the column *Respiratory*. The distributions of quality measures for these highlighted datapoints are shown in Figure 15 using a *Histogram Quality Map*. By comparing Figures 15 and 12, we can see that the highlighted datapoints in Figure 14 have lower quality measures than other datapoints, which means that these values are unreliable.

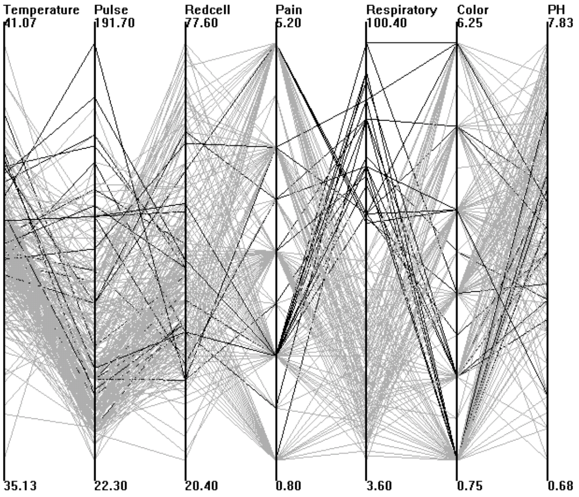


Figure 14. The brush on data space for the dataset *Horse-Colic* to highlight higher values on column *Respiratory* .

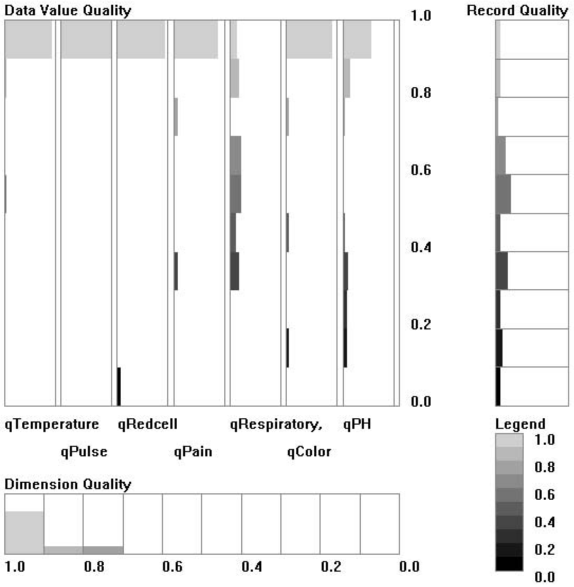


Figure 15. The distribution of quality measures corresponding to datapoints highlighted in Figure 13.

9. Conclusions and Future Work

In this paper, we introduced a novel and useful interaction technique, value-attribute linking, to create a linkage between data space and quality space using brushing. If users select one range in the quality space, the corresponding datapoints in the data space are highlighted. Meanwhile, selected subsets in the data space can result in highlighting the corresponding points in the quality space. We also discussed one special linkage from the quality space to the data space, procedural brushing, using animation with quality measures as the time dimension. In addition, we presented a visualization technique for quality space, *Quality Maps*. It not only can be used in the linkage operations between the two spaces, but also conveys the quality information via a way we feel is more effective than traditional multivariate visualizations.

Some potential future research directions include:

- We plan to explore other mergings of traditional data brushing with quality brushing. For example, whether a datapoint is highlighted could depend on not only its values, but its quality attributes. This might allow users to retrieve more complex patterns. For example, users could focus on a subset with high values on the first data dimension and low record quality.

- Data can be structured in many ways, such as in a hierarchy, and brushing can be done in this structure space [7, 8]. We hope to explore the linking of quality and data space to structure space. In such an extended linkage framework, we can consider performing tasks related to the quality of the data abstraction [5], such as highlighting datapoints based on their data abstraction quality, or examining the quality of sampling or clustering in a selected data subset.

- Other types of data attributes could also be used in addition to quality. For example, data may have access restrictions that could be used as a focus for selection and visualization.

10. Acknowledgments

We gratefully acknowledge our colleagues in the XmdvTool group at WPI for their contributions to this research, and especially Do Quyen Nguyen for the implementation of imputation algorithm to replace missing values. The research funds from NSF and NSA are graciously appreciated.

References

[1] Paul D. Allison. *Missing data*. SAGE Publications, Thousand Oaks CA, 2002.

[2] A. Becker and S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[3] A. Cedilnik and P. Rheingans. Procedural annotation of uncertain information. *Proc. IEEE Symposium on Information Visualization*, pages 77–84, 2000.

[4] H. Chen. Compound brushing. *Proc. IEEE Symposium on Information Visualization*, pages 181–188, 2003.

[5] Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE Visualization and Computer Graphics*, 12(5):709–716, 2006.

[6] DASL. The data and story library [http:// lib.stat.cmu.edu/dasl]. *Cornell University*, 1996.

[7] Y. Fua, M. Ward, and E. Rundensteiner. Navigating hierarchies with structure-based brushes. *Proc. IEEE Symposium on Information Visualization*, pages 58–64, 1999.

[8] Y. Fua, M. Ward, and E. Rundensteiner. Structure-based brushes: A mechanism for navigating hierarchically organized data and information spaces. *IEEE Visualization and Computer Graphics*, 6(2):150–159, 2000.

[9] H. Hofmann and M. Theus. Selection sequences in manet. *Computational Statistics*, 13(1):77–88, 1998.

[10] Shiping Huang. Exploratory visualization of data with variable quality. Master's thesis, Worcester Polytechnic Institute, 2004.

[11] G. J. Hunter. New tools for handling spatial data quality: Moving from academic concepts to practical reality. *URISA Journal*, 11(2):25–34, 1999.

[12] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspective*, 13:10–19, 1992.

[13] A. Martin and M. Ward. High dimensional brushing for interactive exploration of multivariate data. *Proc. IEEE Visualization*, pages 271–278, 1995.

[14] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/mlrepository.html]. *University of California, Irvine, Dept. of Information and Computer Sciences*, 1998.

[15] C. Olston and J. D. Mackinlay. Visualizing data with bounded uncertainty. *Proc. IEEE Symposium on Information Visualization*, pages 37–40, 2002.

[16] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[17] D. Swayne and A. Buja. Missing data in interactive high-dimensional data visualization. *Computational Statistics*, 13(1):15–26, 1998.

[18] Deborah F. Swayne, Andreas Buja, and Duncan Temple Lang. Exploratory visual analysis of graphs in ggobi. *Workshop on Distributed Statistical Computing (DSC 2003),Vienna,Austria*, 2003.

[19] Deborah F. Swayne, Duncan Temple Lang, Andreas Buja, and Dianne Cook. GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43:423–444, 2003.

[20] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, National Institute of Standards and Technology Technical Note 1297, 1994.

[21] J. Tukey, M. Fisherkeller, and J. Friedman. PRIM-9: An interactive multidimensional data display and analysis system. *Dynamic Graphics for Statistics, (W. S. Cleveland and M. E. McGill, eds.)*, pages 111–120, 1988.

[22] A. Unwin, G. Hawkins, H. Hofmann, and B. Siegl. Interactive graphics for data sets with missing values - manet. *Journal of Computaional and Graphical Statistics*, 4(6):113–122, 1996.

[23] P. Velleman. *Learning Data Analysis With Data Desk*. W H Freeman and Co., 1993.

[24] M. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. *Proc. IEEE Visualization*, pages 326–333, 1994.

[25] G. Wills. Selection:524,288 ways to say this is interesting. *Proc. IEEE Symposium on Information Visualization*, pages 54–59, 1996.

[26] C. Wittenbrink, A. Pang, and S. Lodha. Glyphs for visualizing uncertainty in vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):266–279, 1996.

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