Comput Stat (2011) 26:567–584
DOI 10.1007/s00180-010-0226-0

ORIGINAL PAPER

Quality-aware visual data analysis

Matthew Ward · Zaixian Xie · Di Yang · Elke Rundensteiner

Received: 19 September 2008 / Accepted: 30 December 2010 / Published online: 11 January 2011

Abstract The quality, certainty, or confidence of decisions made during the visual analytics process depends on many factors, including the completeness and reliability of the initial data, information loss due to filtering, sampling, and other transformations, and the accuracy and clarity of the visual presentation. Unfortunately, in most visualization tools the analyst is unaware of these and other forces that degrade the meaningfulness of their results. In this paper, we describe our efforts to design strategies for tackling the measurement, display, and utilization of quality aspects at all stages of the visualization pipeline. The goal is to help analysts maintain an awareness of the accuracy and completeness of the information conveyed in the images, and subsequently the patterns observed and decisions made based on the analysis. Quality measures can be used both to assist analysts in selecting, transforming, and mapping their data as well as to automatically refine processes to generate higher quality views. We have implemented several such techniques within XmdvTool, a public-domain package for visual analytics. We illustrate the quality-specific components of our tool with several case studies to show the usefulness of the approach. We also describe user studies that were performed to validate the accuracy of our quality measures.

This work is supported under NSF grants IIS-0119276 and IIS-0414380.

M. Ward (✉) · Z. Xie · D. Yang · E. Rundensteiner
Computer Science Department, Worcester Polytechnic Institute,
Worcester, MA, USA
e-mail: matt@cs.wpi.edu

Z. Xie
e-mail: xiezx@cs.wpi.edu

D. Yang
e-mail: diyang@cs.wpi.edu

E. Rundensteiner
e-mail: rundenst@cs.wpi.edu

Keywords Visual analytics · Quality measures · Information loss

1 Introduction

Exploratory data analysis is emerging as a core technology critical for the success of a large variety of application domains, including homeland security, bioinformatics, finance, and product quality control. Such applications require the exploration and analysis of data sets of often enormous size and complexity. In support of this, visualization takes advantage of the immense power, bandwidth, and pattern recognition ability of human perception and cognition. Visual displays, as well as direct interactions on and with the displayed information, enable analysts to make informed conclusions rapidly yet accurately.

The validity of information extracted from exploratory visualization and the decisions made are, in a large part, dependent on the information available to analysts to draw their conclusions. Clearly, tools to support exploratory data analysis must be of high quality in order to be effective. For visualizations in particular, this implies that information in the images is communicated using the following principles:

- **Accuracy**: The depiction should assign the different pieces of information the appropriate amount of screen space and with appropriate visual mappings to best capture their actual characteristics, i.e., with minimal distortion.
- **Completeness**: The display should contain all information of relevance to the situation being analyzed, i.e., without suffering from information loss.
- **Sufficiency**: The display resolution should be sufficient for the analysis, i.e., playing to the strength and limitations of the human perceptual abilities in terms of being able to discern, for instance, the nuances of the color scale.
- **Intuitiveness**: The visualizations should be readily interpretable to facilitate the focusing of the analyst's perceptual and cognitive skills on the task at hand.
- **Responsiveness**: The display and interaction technology should be responsive to enable analysts to visually explore the information by traversing and manipulating the information space in near real-time; i.e., without undue delays that can interrupt an analyst's train of thought.

In practice, it is extremely difficult to attain all of the above requirements. The reason is that various quality issues exist in the process of information visualization. Clearly, important facets of quality that must be considered in the design of the visual analytics technology include:

- **Certainty**: can I assume (or better yet know) that the underlying data being displayed is indeed accurate, and that the visual mapping conveys that accuracy?
- **Correctness**: how well does the visualization represent the actual data; are there visual artifacts that do not correspond to actual data features?
- **Confidence**: am I confident in the conclusions I have drawn from the visualization, and that they are not the result of misinterpretation?
- **Overload**: how much effort does it take to interpret the additional computation and display of quality measures, and is that overhead overshadowing the benefits gained by providing explicit access to such knowledge?

These challenges for the design of visual analytics tools must be tackled in the context of many facets of quality, including accuracy, completeness, certainty, consistency, or any combination of these. All of these concepts can be studied, computed, and displayed throughout the visual exploratory pipeline. Visual analytics can be regarded as a four-phase pipeline, composed of the stages of data collection, data transformation, graphical mapping and display. Clearly, quality degradation might happen at each phase, as illustrated below:

- **Data collection**: During the acquisition phase, a wide range of error types can be observed. First, some of the data may be missing. In addition, errors or noise might be introduced during the collection or entry of the data due to the quality of the device used for measurement of the data or due to human error in data entry (Pang 2001). Data sets may also be inconsistent, such as a text description appearing in a field where a numeric value is expected (e.g., a date).
- **Transformation**: Data transformation includes processes filtering, smoothing, sampling, clustering, and dimensionality reduction, any of which can cause information to be lost. For example, clustering provides an overview of a large scale dataset to the user, but discards some, possibly important, details.
- **Graphical mapping**: This phase concerns the translation from numerical or nominal values into visual attributes of graphical display entities, such as the thickness of a line or the choice of color of a glyph. The combination of different mapping choices from domain-specific values into graphical attributes must consider the limitation of human perception and avoid the overutilization of graphic attributes with multiple semantics—or risk creating a visualization that is difficult to interpret quickly and accurately.
- **Display**: During this final step of the pipeline, which is responsible for the graphical rendering of the mapping result onto the screen, visual clutter can occur. Clutter refers to visual entities overlapping in the image due to relative positioning and sizing of display entities, or simply the sheer number of entities. This can seriously affect the user's ability to perceive patterns within the visualizations. Note that issues of graphical mapping and display may overlap, such as selecting an appropriate layout.

Clearly, the effectiveness of visual analysis is limited by the visualization itself, that is, the conclusions drawn from the graphic representation are at best as accurate as the visualization. Therefore, to maintain the integrity of visual data exploration it is important to design a visualization so as to convey not only the actual data but also all aspects of its quality (Amar and Stasko 2004). With few exceptions, most current visualization tools have ignored quality issues. They assume that data has been filtered in previous procedures, and treat it as if it were completely reliable and accurate. That is, most current visualization systems do not explicitly convey these important meta-properties about the data. In such visualization tools the analysts are left unaware of these forces that act to degrade the meaningfulness of their results.

In our work, we take a different approach. Our general methodology for tackling quality in the visualization pipeline is to visually convey not only the actual data but also its quality measures explicitly. In addition, we provide users with tools for

improving quality attributes and for trading off between between different quality types, such as information content versus responsiveness.

It is impossible to exhausted discuss all of the issues related to the types of quality mentioned in the above list in depth. Instead, we have identified a core subset of these issues. More precisely, we focus on tackling the following three types of quality:

- **Data quality**: We explicitly address the certainty of the underlying data or lack thereof.
- **Abstraction quality**: We also study how well a given abstraction resulting from the transformations on the data represents the salient features of the original data.
- **Visual quality**: Lastly, we explore the quality of the visual display itself. Clearly, different graphical mappings of the same information may be of varied quality, either bringing out certain features, hiding others, or possibly even causing unwanted artifacts to appear in the display.

Our general methodology is to enable effective exploration by explicitly exposing and integrating the quality features as part of the display itself, thus equipping the human decision maker with insight into and control over the type of data and transformations they work with. This should result in more informed decision-making. Quality measures can be used both to assist analysts in selecting, transforming, and mapping their data as well as to automatically refine processes to generate higher quality views.

For each type of quality, we follow the methodology outlined below to design a quality-aware visualization:

1. Design and implement metrics for measuring this class of quality;
2. Develop customized display techniques to convey this quality to the analyst.
3. Assess measures against the analyst's judgments of perceived quality;
4. Allow analysts to interactively modify aspects of the pipeline stage to enhance quality, trading off between possibly conflicting quality measures; and lastly,
5. Develop automated methods to modify the pipeline stage to enhance the quality, whenever possible.

While the steps of our proposed methodology are general, we validate our proposed ideas by realizing them in concrete tools targeting primarily the analysis of multivariate numeric data and geospatial data. This work was conducted in the context of our NSF supported project on developing a freeware tool suite (XmdvTool) to facilitate interactive data analysis of multivariate data sets (Ward 1994). However, while most of our techniques are equally applicable to other types of data, we leave this for future work. We illustrate the different aspects of our overall quality-centric visualization solution with several case studies to show their usefulness. Our case studies were based on the five core visualization techniques supported within XmdvTool, namely, scatterplot matrices, star glyphs, parallel coordinates, dimensional stacking, and pixel-oriented displays. Clearly, other visualizations could and should equally be augmented to tackle the challenges of measuring and conveying quality for visual analytics. In this manuscript, we also briefly describe user studies that we performed to validate the accuracy of our proposed quality measures.

The remainder of this paper is organized as follows. In Sect. 2 we describe our approach for data quality, including the mappings from quality measures to graphical

variables. Section 3 is dedicated to the discussion of abstraction quality. Section 4 describes our efforts at addressing the visual quality problem. We conclude this paper in Sect. 5 with discussions and possible future research directions.

2 Data quality

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

2.1 Related work

XGobi (Swayne and Buja 1998) and MANET (Missing Are Now Equally Treated) (Unwin et al. 1996; Hofmann and Theus 1998) 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 work regarding data quality issues, focusing on uncertainty definition, modeling, computation and visualization (Hunter 1999; MacEachren 1992; Djurcilov et al. 2002). Many possible graphical variable mappings to represent uncertainty have been proposed, including color, opacity, texture, fog, animation, and flashing. Wittenbrink et al. (1996) and Pang et al. (1997) proposed techniques for visualizing uncertainty found in vector fields. They developed and evaluated many mappings of uncertainty degree to glyph attributes, including adding glyphs, adding geometry, modifying geometry, modifying attributes, animation, sonification, and psycho-visual approaches. Sanyal et al. (2009) performed a user study to compare four commonly used techniques for visualizing uncertainty: errorbars, scaled size of glyphs, color-mapping on glyphs, and color-mapping of uncertainty on the data surface.

All of the above work focus on spatial or temporal univariate data. Although some people have tried to extend the work to other data types, such as multivariate data (Tekušová et al. 2008), the capabilities of these visualization techniques are limited because the techniques only show the quality associated with data records. In order to expand the visual representation of data quality to multivariate data, we must solve three problems: (1) A more complex model to represent data quality is necessary, because each data value, record, and dimension can have their own quality attributes; (2) We should carefully choose mapping methods, taking into consideration perception theory and evaluating the results via user studies; various visual elements are already utilized to represent the data, and mapping quality information onto different graphical attributes can interfere with interpretation; (3) In order to help users explore the enhanced visualizations, interaction techniques are needed.

The first two issues are unique to multivariate data, as compared to univariate data. Thus, we will first present a model for data quality on multivariate data, and then design visualization and interaction techniques for it.

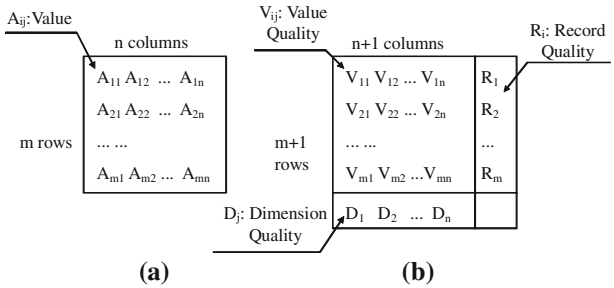

Fig. 1 The structure of data quality defined in this paper. **a** Data space. **b** Quality space

2.2 Proposed quality space

In order to model data quality, we first introduce the notion of **Quality Space**, which is composed of quality measures at three granularities: data value, record, and dimension. As a reasonable starting point, we employed scalar values to measure uncertainty (Xie et al. 2006), on form of data quality. We assumed that quality measures consist of a vector of values for the record quality (one entry per record), a vector for the dimension quality (one entry per dimension), and a two dimensional table of values for the data value quality (one entry per value in the original dataset).

The dimension quality mechanism provides an overall quality measure to the values in each dimension. It keeps the analysts aware of which dimensions are in highly quality and thus can be used as solid evidence for the potential conclusions, while the others may be in low quality and thus less evidential. For example, when an analyst is studying citizens' salary over a census dataset, if she knows that the values in both the *salary* and the *occupation* dimensions are in high quality, she can confidently draw conclusions about the relationship between people's income and their occupations. However, if she knows that dimension quality of *working years* dimension in this dataset is in low quality, she should be aware that any relationship between working years and salary learned from this dataset may be inaccurate.

All values for quality measures are normalized to the range of zero (lowest quality) to one (perfect quality). Figure 1 shows the configuration of these three types of data quality.

Users in various application areas can assign different real meanings to these quality measures. For our research project, we employed a multiple imputation algorithm to generate estimated values for missing values. The value quality measures are computed to represent the degree of reliability for the imputed values, and the record and dimension qualities are averages of the corresponding rows and columns. Other methods for computing quality attributes could equally be used.

2.3 Display

We investigated two alternate approaches for visually conveying the data quality.

(1) **Embedded Display** (Xie et al. 2006): We convey quality measures using graphical attributes of visual elements in existing multivariate visualizations. As we know,

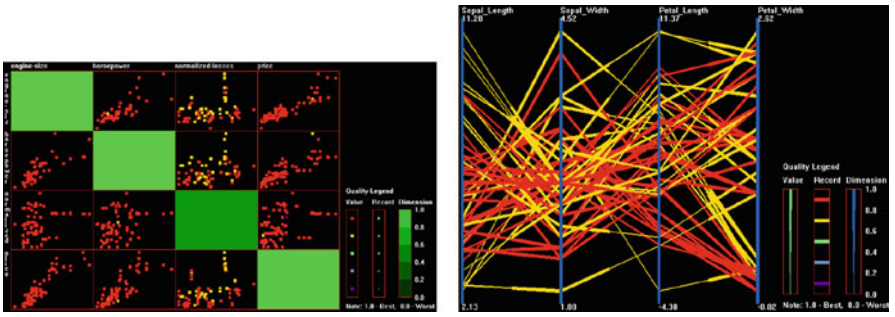

Fig. 2 The embedded display. The quality measures of data values, records, and dimensions are mapped to **1** mark color, mark size and diagonal color; **2** line width, line color and axis color, respectively

normally there exists some unused visual attributes in multivariate visualizations, e.g., mark size in scatterplot matrices and line width in parallel coordinates. Our basic idea is to employ these visual attributes to visually convey quality measures. The three types of quality measure can be mapped to different visual attributes. For example, we can map value quality to line width, and record quality to line color (See Fig. 2). To create an effective embedded display, choosing appropriate mappings is key. Based on perception theory, we identified visual attributes for each multivariate visualization technique that could potentially represent quality information effectively (Xie et al. 2006). One significant problem is that visualizations become overloaded with excessive information when we apply visual encoding to quality measures, because different visual attributes can impact each other, and users may find it difficult to extract multiple visual attributes. In addition, some visual attributes, such as line width and point size, easily causes visual clutter. This consideration led to the development of our second approach.

(2) **Quality Space Display**: This approach provides a view, which we call a *Quality Map*, for data quality, separate from the display of the original multivariate data. Figure 3 shows two types of quality map, stripes and histograms. The former, is adapted from Table Lens (Rao and Card 1994). Each stripe in the *data value quality*, *record quality*, and *dimension quality* sections corresponds to one quality value. The brightness of each stripe reflects the quality measures as shown in the legend at the bottom right corner. For the latter, histograms are introduced to represent the distributions of various types of quality measures. Although these two visualization techniques are not new, the *Quality Map* is a new way of using the techniques. We also enable users to sort records and dimensions based on the quality attributes. Note that, in the left image of Fig. 3, all of quality measures are sorted based on the value quality of the first dimension; thus users can easily see relationships between value quality of different dimensions and record quality.

2.4 Interaction

After we designed the above two types of display, we introduced *value-attribute linking* to help users explore datasets with quality measures (Xie et al. 2007). Two types of linking are supported:

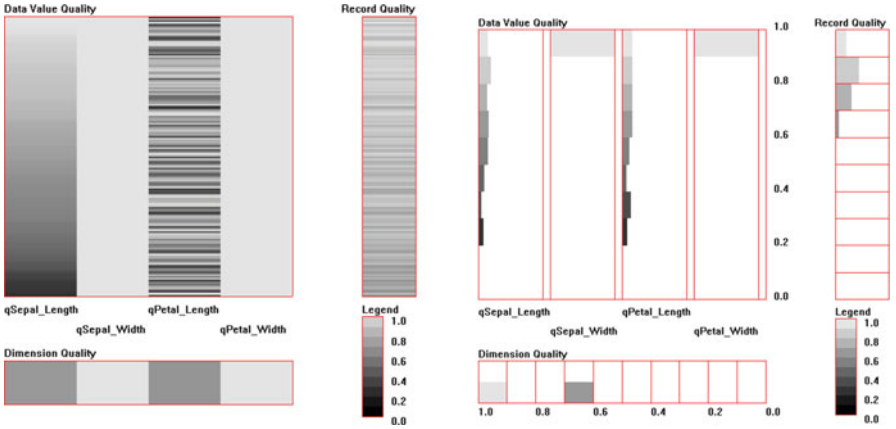

Fig. 3 The separate view for quality space. **1** Stripe quality map: Each *stripe* denotes a quality measure. Records can be sorted based on value quality or record quality. **2** Histogram quality map: we use histograms to represent the distribution of quality measures

Linking from quality space to data space: When users select a range in quality space, all data points falling into this quality range are highlighted in the data space. Note that these data points are not necessarily contiguous in the display. We call this linking *quality brushing*, as compared to N-dimensional brushing (Martin and Ward 1995) and structure-based brushing (Fua et al. 2000). A useful capability of this linking is that it can help the user select data records with high quality. Hence, the user can focus on the data records with high quality to draw reliable conclusions. Note that embedded displays can also highlight data with high quality if we choose appropriate visual attributes. This link operation enables more functionality to explore dataset with variable quality. For example, users can choose a low quality range, with the highlighting in the data space potentially suggesting a reason for the low reliability.

Linking from data space to quality space: When users highlight a subset in data space using traditional N-dimensional brushing, the quality measures of the data-points in this subset are highlighted in the quality space. This linking operation can help users explore the distribution of quality measures for the subset of interest. In addition, this type of linking can confirm some findings from the first type of linking.

2.5 Evaluation

For embedded displays, a user study was carried out to attempt to determine the visual variables in parallel coordinates, scatterplot matrices, and star glyphs that can convey quality information most effectively. In this study, the visual attributes tested included line width, brightness, and hue for parallel coordinates/star glyphs, and dot size, brightness, and hue for scatterplot matrices. First we created some artificial datasets having quality measures, and then used one visual attribute to convey the data quality within the normal visualization. Participants were asked to perform some quality-related tasks on the final visualization, e.g., focusing on high-quality data and trying to draw

reliable conclusions. Finally, the response accuracy and response time of subjects were collected and analyzed. From this user study, we note two important conclusions: (1) No visual attribute was consistently good across all visualizations. For example, hue performs very well in parallel coordinates, while point size is the best for scatterplot matrices when datasets are small. (2) When datasets become large, all of visual attributes generated poor results. For this reason we felt it necessary to design a separate view for the quality space.

2.6 Future research opportunities

In order to incorporate data quality visualization in complex applications and provide an approach to making the retrieval of quality-related patterns easier and quicker, there are many possibilities for future work: (1) For some applications, we need a more complex data quality model. For example, bounded uncertainty (Olston and Mackinlay 2002) gives a precise lower and upper bounds to convey the uncertainty, so that a vector is needed to represent a quality measure. Currently, our proposed model only adds an extra numerical quality measure to each data value, record, and dimension. (2) Additional user studies should be conducted to study other mappings of quality variables to visual attributes. One potential experimental goal would be to test how two visual variables affect each other, such as using line width for data value quality and line color for record quality. (3) The types of visualizations extended with quality information should be expanded to include other multivariate visualization techniques as well as scientific data and graph visualizations.

3 Abstraction quality

In this section, we discuss how to measure and visualize the amount of information lost during transformations, in particular, the abstraction process.

3.1 Related work

Data abstraction is the process of reducing a large dataset into one of moderate size, reducing the level of detail while maintaining dominant characteristics of the original dataset. Different techniques have been proposed to abstract data and thus address the scalability problem for visualization and other data analysis procedures, including sampling (Dix and Ellis 2002), clustering (Fua et al. 2000), filtering (Ahlberg and Shneiderman 1994) and dimensionality reduction (Carreira-Perpinan 1997). However, few efforts have been reported on measuring and conveying information lost during these and other transformations. Bertini and Santucci (2004) presented a quality measure for sampling and applied it to finding the optimal sampling level. This measure, however, was limited to sampling. The authors did not consider other types of abstraction. Boutin and Hascoet (2004) reviewed and compared various measures for graph clustering, and proposed new methods. These measures were designed specifically for clustering and their extensibility into other types of transformations is not clear.

Luo et al. (2003) presented techniques to compare histograms formed based on different datasets. However, the techniques proposed in this work were not specifically designed for capturing abstraction quality. Wang and Ma (2008) proposed a reduced-reference approach to assess quality loss in the reduced or distorted version of volume data. Because this approach was specially designed for volume data, it is not easy to extend it to other types of data.

3.2 Proposed abstraction quality measures

Our proposed abstraction quality measures are designed to measure how well the transformed data (after abstraction) represents the original data (before abstraction). Our abstraction quality measures capture the level of "preservation" from the original dataset to an abstracted dataset that has the same semantics, such as a multi-variate dataset D and a multi-variate dataset $D_{abstract}$ abstracted from D. Such measures are essential for effective data analysis; with them, researchers can make informed decisions based on the transformed data, as well as choose between alternate methods for data processing. In particular, we have employed three abstraction quality measures, namely Statistical Measure (**SM**), Histogram Difference Measure (**HDM**) and Nearest Neighbor Measure (**NNM**).

Many statistical measures can be used to represent and compare datasets. Perhaps the most commonly used statistics are mean value and standard deviation. For our experiments we used differences in mean values as one way of measuring data abstraction quality. The **SM** (Statistical Measure) is defined as the distance between two sets of statistics. In particular, we first compute the mean value of each dimension for the original dataset and the abstracted dataset, respectively. Then, we compute the distance between mean values of the original dataset and the abstracted dataset by summing their absolute difference on all dimensions. The result is one measure of Data Abstraction Quality (DAQ).

As a histogram is a common data descriptor and is fast to compute, we use the difference between the normalized histograms of the original dataset and the abstracted dataset as another measure of the DAQ. First, we compute two histograms with the same number of bins from the original dataset and the abstracted dataset. If the distributions are skewed, we can use non-uniform bin widths. Bin difference is defined as the absolute difference between two bins. Then the histogram difference corresponds to the summation of bin differences between the corresponding bins in the two histograms. The **HDM** (Histogram Difference Measure) is defined as the normalized histogram difference. Its range is from 0 to 1. 0 means in every pair of corresponding bins, at least one is empty, and 1 indicates a perfect match. In particular, the HDM for a single dimension is calculated by averaging the difference between the percentages of objects falling into each bin on this dimension. Finally, we can calculate the abstraction quality by summing the HDM for all dimensions.

Our third DAQ is based on a nearest neighbor algorithm. As the name implies, nearest neighbor algorithms (Duda et al. 2001) search for the object nearest to a given object. They are widely used to classify data into groups in data clustering and pattern recognition. Every object corresponds to a record. We assume that each record in the original dataset has a nearest neighbor in the abstracted dataset, called

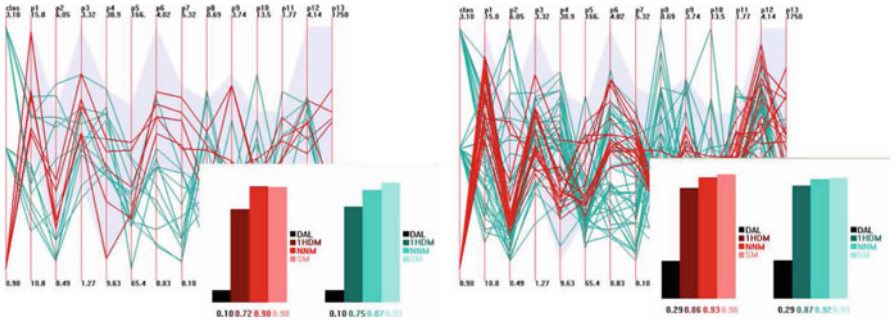

Fig. 4 Two different datasets abstracted from the wine dataset (Asuncion and Newman 2007). The *bar chart* on the *right lower corner* of each abstracted dataset shows the abstraction quality calculated by different measures for the selected and unselected data

its representative. The records in the original dataset that are represented by the same record in the abstracted dataset form a cluster. We define the **NNM** (Nearest Neighbor Measure) as the normalized average of distances between every record in the original dataset and its representative. More details of all three measures introduced above can be found in Cui et al. (2006).

3.3 Display

Once we have the abstraction quality calculated, different visualization techniques can be applied to display it. We now describe two possible methods for displaying the abstraction quality. The first is general purpose, and can be applied to display the abstraction quality calculated for any abstracted dataset. The second is specifically designed to display the abstraction quality for data clustering techniques. Since many multivariate visualization systems support interactive selection via brushing (Martin and Ward 1995) using a rich assortment of tools, our first display method visualizes the multiple measures using separate sets of bar charts for the selected and unselected data, respectively. Figure 4 gives an example of the abstraction quality calculated by different measures for both selected (red bars) and unselected (green bars) data in two abstracted datasets.

The second display method visualizes the abstraction quality of a hierarchically clustered dataset using InterRing (Yang et al. 2002), a radial, space-filling hierarchy visualization method. As shown in Fig. 5, the rings represent layers of the cluster hierarchy, with the root node in the center. Each arc in a ring represents a cluster in the corresponding level. Deeper nodes of the hierarchy are drawn further from the center, and child nodes are drawn within the arc subtended by their parents. The sweep angle of a leaf node is proportional to the cluster radius, and the sweep angle of a non-leaf node is the aggregation of all its children. Color is used to convey the quality of each cluster, where red is low quality and blue is high.

3.4 Interaction

Beyond the display of abstraction quality, we also introduce interaction techniques to help users select the appropriate data abstraction level. On one hand, users can adjust

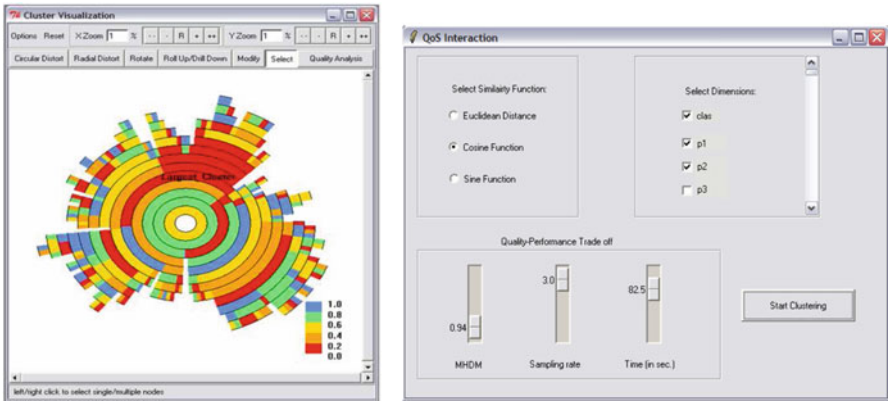

Fig. 5 *Left*: A cluster hierarchy built for the wine dataset (Asuncion and Newman 2007) and displayed by InterRing. *Right*: The control panel for quality-aware abstraction

parameters of the abstraction transforms; new measures will then be generated and displayed. For the bar chart display method, users can move a slider bar to adjust the Data Abstraction Level (DAL) of a dataset. After the DAL has been changed, the system will generate an abstracted dataset and display it in the data visualization. Also, the new abstraction quality will be calculated for this abstracted dataset and displayed in the bar chart. The DALs for selected and unselected data can be adjusted independently. On the other hand, users can modify the location of one of the boundaries of the selected region by clicking the left mouse button on or near the boundary and dragging in the desired direction. In addition, the selected region can be moved by choosing a region on the data display and then adjusting the DAL for the region. For the InterRing display, users can directly operate on the InterRing to improve the quality of clustering via splits and merges. More specifically, a user can merge any two clusters in the same level by dragging one to another, and she can split any cluster by double clicking the target cluster with the mouse.

Besides direct control on the transforms, users can also adjust the abstraction quality measure display and thus indirectly control the transforms by using the control panel shown in Fig. 5 (right part). In general, a user can customize the abstraction process by trading off between abstraction quality, system resource utilization, and the time needed for computing the abstraction. Other abstraction related settings, such as the distance function for data records, can also be customized from this panel. Such controls are especially important when the cost of transforms is expensive, indicating that generating a high quality abstraction may require significant system resources, long processing times, or both. In particular, a user can pre-set a target abstraction quality before the transform. Then the system will feedback the estimated processing time and system resources needed to finish the abstraction. If the user is satisfied with the estimation, our system will adjust the transform process to fulfill the user's requirements, indicating that the abstracted dataset generated will have the abstraction quality as specified. If the user cannot afford the long processing time or the system resources as estimated, she can re-set the target abstraction quality and the system will redo the estimation.

3.5 Evaluation

We conducted an evaluation to check how well the data abstraction measures conform to the data abstraction quality perceived by users. First, we generated multiple abstractions (samples, clusters) for several data sets, and then asked subjects to rate the degree to which the abstraction represents the original data. Finally we compared the users' assessments with the abstraction quality calculated by different measures.

After careful comparison of the quality level of each measure with users' estimates, we identified the following variants on each measure that best matched user assessment: HDM using an N-D histogram and Manhattan distance, NNM using a customized normalization method (distance between records are normalized based on the distance between two records that are farthest away to each other in the dataset)and Euclidean distance, and SM using mean values and Euclidean distance. Also, we found that different abstraction measures may be sensitive to the changes in different dataset features, such as the relative data density and characteristics of outliers. In particular, HDM is more sensitive to changes in relative density caused by the abstraction, while NNM is more sensitive to the outliers in the dataset. We concluded that having multiple measures shown for each abstracted dataset is a better strategy than just choosing one, as the analysts can better assess the abstraction quality of a dataset from multiple perspectives.

3.6 Future research opportunities

Many potential research opportunities are still open for future study on the topic of data abstraction quality. First, new abstraction quality measures could be developed that more accurately predict users' estimations for the specific abstraction methods. Second, there is a scalability problem, namely how to quickly calculate the abstraction quality when both the number of records and dimensions grow dramatically. Third, the current quality displays are separated from the data display, which may divert an analyst's attention. New display methods could be designed to better integrate the quality displays with data displays. Fourth, more studies need to be conducted to discover the strengths and weaknesses of each abstraction quality measure in conveying different information, i.e., which one is more sensitive to different data features, such as trends and outliers.

4 Visual quality

In general, there are multitudes of ways to visualize a given data set; not only can the data be reconfigured (e.g., through reordering of records or dimensions), but also the specific graphical mappings and viewing parameters can be changed. The question we pose is whether there are metrics that can help evaluate the quality of a visualization, i.e., how well does it convey information? These metrics, if validated, could be used to both identify subsets of effective visualizations out of a potentially large pool as well as to refine visualizations to be more readily interpreted.

4.1 Related work

The quality of a visualization can be measured in many ways. Several researchers have proposed techniques in the past. Haase (1998) proposed six categories of evaluation criteria: data resolution, semantic quality, mapping quality, image quality, presentation/interaction quality, and multi-user quality. Each tool or technique could be evaluated subjectively on each of these attributes. However, upon reflection we would claim these are measures of the *power* or *flexibility* of a system, and not the type of quality we hope to be able to accurately measure. Friendly and Kwan (2003) introduced the term *effect ordering* as a mechanism to reorder the elements in a visualization to emphasize particular patterns. They showed how eigenvalue and singular value decomposition methods could be used in techniques such as parallel coordinates and star glyphs. While related to some of our work, they did not explicitly work towards the removal of visual clutter. Rosenholtz et al. (2005) use color and luminance contrast to measure *feature congestion* in a visualization. Tests using human observers ranking sets of visualizations confirmed these measures could be used to predict the responses of human subjects. A major difference between our work and theirs is that our measures are customized to each type of visualization, as we believe the notion of visual clutter is tightly dependent on the specific graphical mappings used.

4.2 Proposed visual quality measures

One of our conjectures is that images with high levels of visual clutter are, in general, less useful than images with low levels of clutter. Similarly, images with high levels of visual structure are more useful than those with low structure. Our approach (Peng et al. 2004), therefore, has been to design and develop both clutter and structure measures for different visualizations and use these measures to identify visualizations with the lowest clutter or highest structure. Below we briefly describe the measures used (details can be found in Peng et al. (2004)):

- Scatterplot matrices: plots with similar characteristics should be near to each other. We first separated plots involving discrete variables (low cardinality) from those involving two continuous variables (high cardinality). For the former, we used the difference in the number of discrete values for the variables in neighboring plots. For the latter, we compared the Pearson's correlation coefficient between adjacent plots (other measures could work as well, such as the scagnostics measures presented in Wilkinson et al. (2005)).
- Parallel coordinates: outliers between adjacent axes should be minimized. Outliers were determined using a 2-D nearest neighbor search; if the nearest neighbor is further than a particular threshold, the point was considered an outlier. Thus a point that is an outlier only in one of the dimensions would, in general, be more isolated than an outlier in 2 adjacent dimensions.
- Star glyphs (or other shape-based glyphs): shapes should have low complexity (measured by the number of concavities) and be as symmetric as possible. The conjecture (yet to be validated) was that simple shapes are easier to remember than complex ones, and that having most of the shapes be simple and symmetric makes

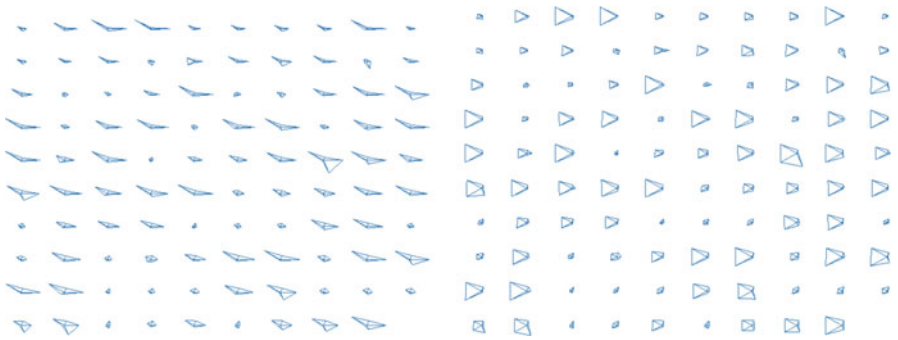

Fig. 6 The coal mining disaster dataset displayed with star glyphs. The image on the *left* is drawn using a random dimension order (clutter score = 230), while the one on the *right* uses an optimized ordering (clutter score = 54)

finding patterns easier. The overall clutter was proportional to the percentage of glyphs that were considered complex or the sum of the individual clutter measures.
– Dimensional stacking (or other matrix-based visualizations): the number of distinct components or groups of adjacent occupied cells should be minimized. Our conjecture was that a small number of adjacent islands of data would be easier to analyze than plots with significant scattering of the data.

4.3 Proposed clutter reduction methods

Clutter reduction methods can fall into three distinct categories: information-preserving, information-lossy, and remapping. For this research effort, we focused on information-preserving methods. The variable we chose to adjust was dimension ordering, as this does not result in data being discarded or deemphasized, yet can have a significant impact on the levels of clutter or structure in many visualizations. We implemented both exhaustive search techniques (looking at all possible orderings) as well as heuristic approaches based on hill-climbing strategies to find local minima of the clutter measure (or maxima of the structure measure). Figure 6 shows an example of a visualization before and after dimension order optimization.

4.4 Evaluation

In order to validate that our visual clutter and structure measures were reasonable, we performed a user study involving 13 subjects; 5 were visualization experts and the rest were novices. After training on the interpretation of each of the visualization techniques, they were presented with 24 pairs of visualizations. Each pair consisted of a particular data set presented with one of the four visualization techniques mentioned above: one with the original dimension ordering and the other with the optimized ordering (the ordering of each pair was randomized). For each, the subject was asked to choose one of three assessments: A is less cluttered/more structured than B, B is less cluttered/more structured than A, or they are roughly equivalent (no preference).

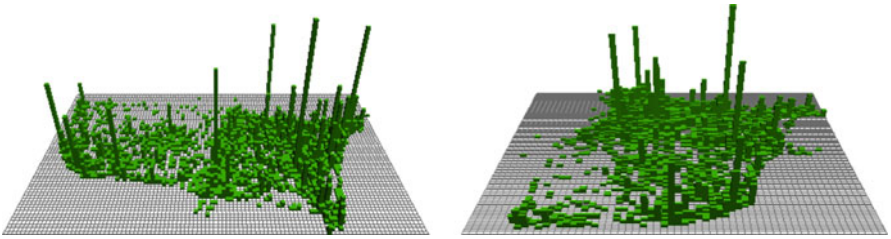

Fig. 7 A cityscape view of air traffic over the United States. The view on the *left* has a high degree of occlusion, especially along the East Coast, which is significantly reduced in the view on the *right*

Subjects were given 15 seconds per image pair. For each type of visualization, between 4 and 8 different data sets were tested.

For expert users, in 22 of the 24 cases either they found the views to be roughly equivalent, or the optimized order was preferred. For the other 2 cases, the non-optimized order was preferred in one case, and in the other, each ordering received the same number of votes. In many cases, the optimized ordering was preferred by all of the subjects. For the non-expert users, the results were somewhat less consistent. In 17 of the 24 cases, they preferred the optimized order, in 6 cases, the non-optimized order, and in 1 case, they were evenly split. We attribute this difference to the fact that this group of subjects had minimal experience in doing visual data analysis using these particular visualization methods. Still, we believe that the results indicate that our measures have some validity and utility in finding dimension orders that were perceived as generating less cluttered displays.

Note that these evaluations did not test issues of retention or task speed/accuracy, only perceived visual clutter. Further testing is needed to confirm that increased visual clutter results in poorer task performance.

4.5 Future research opportunities

Clearly, there are many potential avenues for continued research into visual quality measurement and use. For example, alternate measures of visual clutter and/or structure are possible, and should be tested. We are also investigating measures for other types of visualizations. In one such effort, we tested the concept of optimizing camera positions in 3-D cityscape visualizations, using measures of occlusion to find the views that minimized information loss due to occlusion (see Fig. 7). Also, most of our efforts have focused on information-preserving techniques. There is much work that can be done in information-lossy and remapping strategies for visual clutter reduction.

5 Conclusions and future work

In this paper, we addressed a limitation found in practically all visualization tools, namely that they leave the analyst unaware of the forces that degrade the meaningfulness and certainty of their results. In this paper, we described our efforts to create tools for the explicit measurement, display, and utilization of quality aspects at all

stages of the visualization pipeline. The goal is to help analysts maintain an awareness of the accuracy and completeness of the information conveyed in the images, and subsequently the patterns observed and decisions made based on the analysis. One key principle we advocate in this work is that to move data and information visualization from ad-hoc development of tools to a science, we need some measures of goodness or quality that can be used for both quantitative and qualitative assessment. We have provided arguments as well as case studies to support that quality can be measured at all stages of the visualization pipeline, from the data extraction to its visual presentation on the screen. Quality measures can be used to assist analysts in selecting, transforming, and mapping their data as well as to automatically refine processes to generate higher quality views. Given a sense of the quality of what is being shown and the quality of how it is being shown, analysts can associate a confidence level with their observations and decisions. Our preliminary case studies confirm the usefulness of this approach.

Our effort into providing rigorous treatment of all aspects of quality within the visual exploration pipeline, while one step forward, opens numerous avenues for further study.

- Clearly, many alternate measures for the different types of quality are possible, and thus should be explored.
- Once new quality metrics are being considered, then a rigorous testing using both user studies and case studies should be undertaken to assess those newly proposed metrics.
- While we outline our initial steps in measuring and conveying quality in interactive visual data exploration, in the long-term future we envision that such information should become an integral part of any and all data visualizations.
- The holy grail is the ability to directly associate the measured quality with the confidence of final decisions derived while performing visual analysis.

References

Ahlberg C, Shneiderman B (1994) Visual information seeking using the filmfinder. In: Proceedings of the ACM SIGCHI conference on human factors in computing systems, 2:433

Amar R, Stasko J (2004) A knowledge task-based framework for design and evaluation of information visualizations. In: Proceedings of the IEEE symposium on information visualization, pp 143–150

Asuncion A, Newman D (2007) UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

Bertini E, Santucci G (2004) Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In: Proceedings of 4th international symposium on smartGraphics, pp 77–89

Boutin F, Hascoet M (2004) Cluster validity indices for graph partitioning. In: Eighth international conference on information visualisation (IV'04), pp 376–381

Carreira-Perpinan M (1997) A review of dimension reduction techniques. Tech. Rep. CS–96–09, Dept. of Computer Science, University of Sheffield, http://faculty.ucmerced.edu/mcarreira-perpinan/papers/cs-96-09.pdf

Cui Q, Ward M, Rundensteiner E, Yang J (2006) Measuring data abstraction quality in multiresolution visualizations. IEEE Trans Vis Comput Graph 12(5):709–716

Dix A, Ellis G (2002) By chance—enhancing interaction with large data sets through statistical sampling. In: Proceedings of advanced visual interfaces, pp 167–176

Djurcilov S, Kim K, Lermusiaux P, Pang A (2002) Visualizing scalar volumetric data with uncertainty. Comput Graph 26(2):239–248

Duda R, Hart P, Stork D (2001) Pattern classification. 2nd edn. Wiley, London
Friendly M, Kwan E (2003) Effect ordering for data displays. Comput Stat Data Anal 43:509–539
Fua Y, Ward M, Rundensteiner E (2000) Structure-based brushes: a mechanism for navigating hierarchically organized data and information spaces. IEEE Trans Vis Comput Graph 6(2):150–159
Haase H (1998) Mirror, mirror on the wall, who has the best visualization of all? a reference model for visualization quality. In: Proceedings of visualization in scientific computing '98, pp 117–128
Hofmann H, Theus M (1998) Selection sequences in MANET. Comput Stat 13(1):77–88
Hunter G (1999) New tools for handling spatial data quality: moving from academic concepts to practical reality. URISA J 11(2):25–34
Luo A, Kao D, Pang A (2003) Visualizing spatial distribution data sets. In: VISSYM '03: proceedings of the symposium on data visualisation 2003, Eurographics Association, Aire-la-Ville, Switzerland, pp 29–38
MacEachren AM (1992) Visualizing uncertain information. Cartogr Perspect 13:10–19
Martin A, Ward M (1995) High dimensional brushing for interactive exploration of multivariate data. In: Proceedings of the IEEE visualization, pp 271–278
Olston C, Mackinlay J (2002) Visualizing data with bounded uncertainty. In: Proceedings of the IEEE symposium on information visualization, pp 37–40
Pang A (2001) Visualizing uncertainty in geo-spatial data. In: Proceedings of worshop on the intersections between geospatial information and information technology
Pang A, Wittenbrink C, Lodha S (1997) Approaches to uncertainty visualization. Vis Comput 13(8): 370–390
Peng W, Ward M, Rundensteiner E (2004) Clutter reduction in multi-dimensional data visualization using dimension reordering. In: Proceedings of the IEEE symposium on information visualization, pp 89–96
Rao R, Card S (1994) The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: Proceedings of ACM SIGCHI conference on human factors in computing systems, pp 318–322
Rosenholtz R, Li Y, Mansfield J, Jin Z (2005) Feature congestion: a measure of display clutter. In: CHI '05: Proceedings of the SIGCHI conference on human factors in computing systems, pp 761–770
Sanyal J, Zhang S, Bhattacharya G, Amburn P, Moorhead R (2009) A user study to compare four uncertainty visualization methods for 1D and 2D datasets. IEEE Trans Vis Comput Graph 15(6):1209–1218
Swayne D, Buja A (1998) Missing data in interactive high-dimensional data visualization. Comput Stat 13(1):15–26
Tekušová TT, Knuth M, Schreck T, Kohlhammer J (2008) Data quality visualization for multivariate hierarchic data. In: InfoVis demo. http://www.gris.informatik.tu-darmstadt.de/~tschreck/papers/infovis08-poster.pdf
Unwin A, Hawkins G, Hofmann H, Siegl B (1996) Interactive graphics for data sets with missing values—MANET. J Comput Graph Stat 5(2):113–122
Wang C, Ma K (2008) A statistical approach to volume data quality assessment. IEEE Trans Vis Comput Graph 14(3):590–602
Ward M (1994) Xmdvtool: Integrating multiple methods for visualizing multivariate data. In: Proceedings of the IEEE visualization, pp 326–333
Wilkinson L, Anand A, Grossman RL (2005) Graph-theoretic scagnostics. In: Proceedings of the IEEE symposium on information visualization, pp 157–164
Wittenbrink C, Pang A, Lodha S (1996) Glyphs for visualizing uncertainty in vector fields. IEEE Trans Vis Comput Graph 2(3):266–279
Xie Z, Huang S, Ward MO, Rundensteiner EA (2006) Exploratory visualization of multivariate data with variable quality. In: Proceedings of the IEEE symposium on visual analytics science and technology, pp 183–190
Xie Z, Ward MO, Rundensteiner EA, Huang S (2007) Integrating data and quality space interactions in exploratory visualizations. In: Proceedings of the 5th international conference on coordinated & multiple views in exploratory visualization, pp 47–60
Yang J, Ward M, Rundensteiner E (2002) InterRing: an interactive tool for visually navigating and manipulating hierarchical structures. In: Proceedings of the IEEE symposium on information visualization, pp 77–84