A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets

Jibonananda Sanyal, *Student Member, IEEE*, Song Zhang, *Member, IEEE*, Gargi Bhattacharya, Phil Amburn, *Member, IEEE*, and Robert J. Moorhead, *Senior Member, IEEE*

Abstract—Many techniques have been proposed to show uncertainty in data visualizations. However, very little is known about their effectiveness in conveying meaningful information. In this paper, we present a user study that evaluates the perception of uncertainty amongst four of the most commonly used techniques for visualizing uncertainty in one-dimensional and two-dimensional data. The techniques evaluated are traditional errorbars, scaled size of glyphs, color-mapping on glyphs, and color-mapping of uncertainty on the data surface. The study uses generated data that was designed to represent the systematic and random uncertainty components. Twenty-seven users performed two types of search tasks and two types of counting tasks on 1D and 2D datasets. The search tasks involved finding data points that were least or most uncertain. The counting tasks involved counting data features or uncertainty features. A 4×4 full-factorial ANOVA indicated a significant interaction between the techniques used and the type of tasks assigned for both datasets indicating that differences in performance between the four techniques depended on the type of task performed. Several one-way ANOVAs were computed to explore the simple main effects. Bonferronni's correction was used to control for the family-wise error rate for alpha-inflation. Although we did not find a consistent order among the four techniques for all the tasks, there are several findings from the study that we think are useful for uncertainty visualization design. We found a significant difference in user performance between searching for locations of high and searching for locations of low uncertainty. Errorbars consistently underperformed throughout the experiment. Scaling the size of glyphs and color-mapping of the surface performed reasonably well. The efficiency of most of these techniques were highly dependent on the tasks performed. We believe that these findings can be used in future uncertainty visualization design. In addition, the framework developed in this user study presents a structured approach to evaluate uncertainty visualization techniques, as well as provides a basis for future research in uncertainty visualization.

Index Terms—User study, uncertainty visualization.

---◆---

1 INTRODUCTION

Insight and quality assurance can be improved by recording uncertainty along with data. The practical benefits of understanding uncertainty in a scientific process can be manifold. A cursory glance at literature in almost all disciplines will indicate that visual representations to depict the recorded or calculated uncertainties are underdeveloped. The incidence of finding charts of one-dimensional data augmented with errorbars is reasonably high; however, as we move on to data of higher dimensions, visual metaphors to represent the uncertainty are rarely used. Part of the reason is that although a variety of techniques have been suggested, successfully applying them to make insightful visualizations is a very challenging task and there is a lack of guidance on which uncertainty method will yield the best results. In this paper, we present our findings from an uncertainty visualization user study which we believe could help future visualization designs.

Uncertainty can be introduced to the data at various stages in the uncertainty visualization pipeline. Pang et al. [31] divided the pipeline into three stages: data acquisition, data refining, and visualization, and showed how uncertainty can be introduced in any or all of these stages. It is worthwhile to note that uncertainty is an inherent part of all data and designing an uncertainty representation is highly dependent on the type of data being visualized. From our communication with scientists and engineers in various disciplines, researchers agreed that data visualization with an uncertainty representation is highly desirable; however, it is also important to ensure that the user is always aware of the underlying data.

Some uncertainty visualization techniques seem to appear more effective than others, however, little comparison has been done to evaluate the effectiveness of most of these techniques. Keeping this in mind, we categorize uncertainty visualization techniques for one-dimensional, two-dimensional, three-dimensional and temporal data. Based on this categorization, we constructed a user study to evaluate the effectiveness of four commonly used uncertainty visualization techniques:

- Size of the uncertainty glyphs
- Color of the uncertainty glyphs
- Color of the data surface
- Errorbars

In this study, these techniques were applied to both one dimensional (1D) and two-dimensional (2D) simulated datasets. In this paper, we define 1D data as samples from a curve, which is a 1D manifold embedded in a 2D Euclidean space. We define 2D data as samples from a surface, which is a 2D manifold embedded in a 3D Euclidean space. Moreover, the definition for our datasets is such that the 1D data is defined by a 1D function $f(x)$, and the 2D data defined by a function $f(x,y)$. These definitions of 1D and 2D data in this paper were chosen keeping applications of geoscience visualization and analysis in mind.

We asked our participants two types of search questions and two types of counting questions. The following sections discuss the background of the study, our study, and our results in more detail.

- *Jibonananda Sanyal, Phil Amburn and Robert J. Moorhead are with the Geosystems Research Institute, High Performance Computing Collaboratory, Mississippi State University, E-Mail: {jibo, amburn, rjm}@gri.msstate.edu.*
- *Song Zhang is with the Computer Science and Engineering Department, Mississippi State University, E-Mail: szhang@cse.msstate.edu.*
- *Gargi Bhattacharya is with the Department of Psychology, Southern Illinois University-Carbondale, E-Mail:gargi@siu.edu.*

2 BACKGROUND

For a long time, the term uncertainty was used in a rather loose sense. Researchers recognized the need to clearly define uncertainty as it was a necessary step before trying to solve the visualization problem [3, 14, 31]. The International Committee of Weights and Measures (CIPM) described two classes of uncertainty based on the methods used in estimation [5], one of which is derived by statistical methods (such as standard deviations and Analysis of Variance), called the Type A evaluation of standard uncertainty, and the other which is based on scientific judgement (such as experience and specifications), called the Type B evaluation of standard uncertainty. They also defined uncertainty to have 'random' and 'systematic' components which are conditioned by a mathematical model.

The National Institute for Science and Technology (NIST) [36] suggested a 'combined standard uncertainty'. A law of the propagation of uncertainty was given which was based on a root-of-the-sum-of-the-squares (RSS) method. 'Expanded uncertainty' was a term coined to express uncertainty defined by an interval that bounded the measurement. Their definition included correction factors arising from recognized system effects.

A number of techniques have been proposed to study and classify uncertainty visualizations. Johnson and Sanderson [16] provide an overview of the current research and identify important goals for further research. MacEachren [25] touched upon a number of aspects of geospatial uncertainty such as understanding the difference between data quality and uncertainty, and application of Bertin's graphic variables [1]. He discussed conceptual models of spatial uncertainty in a cartographic context and stressed the need to understand the objective of a visualization to justify a good design.

Pang et al. [31] broadly classified the techniques into adding glyphs, adding geometry, modifying geometry, modifying attributes, animation, sonification, and psycho-visual approaches. Recently, Thomson et al. [37] presented a typology to visualize uncertainty information pertinent to geospatially referenced data. Their typology was developed keeping the tasks of an information analyst in mind. Gershon [9] presented a short discussion on imperfections in information, and a taxonomy of the causes of imperfection in knowledge, stressing the need to develop better representations. Some researchers have also discussed uncertainty cataloguing techniques [6, 23].

Olston and Mackinlay [30] argued that visualization methods should be different for statistical uncertainty and bounded uncertainty, since statistical uncertainty representations potentially incorporate infinite ranges of values. They proposed ambiguation as the solution in which statistical graphics are modified or augmented with visual cues that enhance the notion of unboundedness. Huang [13] showed how a multivariate scatterplot can be created by overloading the visual channels such as color, size and background color to show the quality of information.

Efforts have been made to identify potential visual attributes that could be used for uncertainty visualization. Hengl and Toomanian [12] illustrated how color mixing and pixel mixing can be used to visualize uncertainty arising from prediction error in spatial prediction models from soil science applications. Jiang et al. [15] used hue and lightness to show fuzzy spatial datasets. Davis and Keller [8] identified value, color, and texture as potentially the best choices for representing uncertainty on static maps. More recently, Hengl [11], like MacEachren [25], made a case for using hue, saturation, and intensity, suggesting the inverse mapping of color saturation to the magnitude of uncertainty.

A number of disparate fields of research have successfully researched and applied uncertainty visualization techniques. Schmidt et al. [33] looked at ways of representing the multivariate nature of bathymetric uncertainty. Rheingans and Joshi [32] visualized the positional uncertainty of molecules. Strothotte et al. [35] used non-photorealistic techniques to present uncertainty of architectural reconstructions. Li et al. [21] visualized uncertainty in astrophysical data and Lundstrom et al. [24] presented a probabilistic animation method to illustrate uncertainty in medical volume renderings. Lodha et al. [22] used sound for the depiction of uncertainty and Cedilnik and Rheingans [4] demonstrated procedural annotation techniques.

Harrower [10] asked a more fundamental question as to whether the presentation of uncertainty on maps alters the way people solve problems and emphasizes the need to conduct longitudinal studies to identify reasons why a
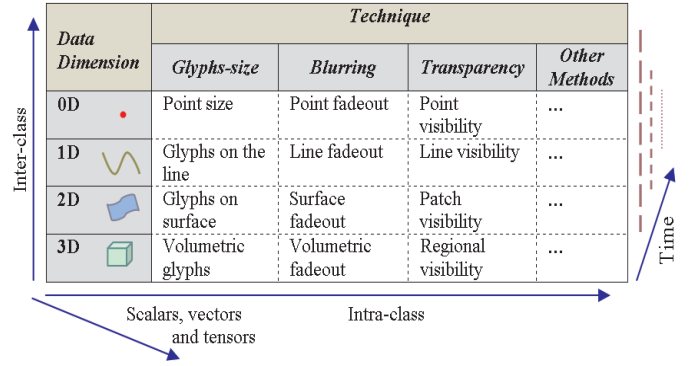

Fig. 1. Our Uncertainty Visualization Framework

subject made a correct or an incorrect choice given an uncertainty representation.

Zuk and Carpendale [42] presented a theoretical analysis of uncertainty visualization in which they evaluated eight uncertainty visualizations from various sources on widely accepted visualization principles from Bertin [1], Tufte [39] and Ware [40]. They presented a set of heuristics and how pertinent each heuristic was with respect to the sampled visualizations. They stressed the need for more research in human factors and perception.

There have been some user studies in understanding effectiveness of uncertainty visualization but most were domain specific studies [17, 19, 26, 34, 41]. There is some justification to that. Uncertainty representations in many processes are inherent and unique to the task at hand. Blenkinsop and Fisher [2] conducted a user study to evaluate uncertainty visualizations of fuzzy classification of satellite imagery. They found that users were highly successful at determining classification uncertainty using greyscale representations in comparison to random animation and serial animation. They found that serial animation performed the weakest.

Leitner and Buttenfield [19] conducted an experiment where participants had to make two sitting decisions, one at a park followed by another at an airport based on a set of predetermined planning criteria. The authors found that addition of certainty information significantly improved the number of correct responses. Additionally, they found that color saturation was not especially effective.

Our objective was to design a systematic and general user study to evaluate the effectiveness of common uncertainty visualization techniques. We generated synthetic 1D and 2D datasets to avoid being tied to any specific application domain. We also chose common tasks in scientific data analysis, such as searching and counting, to evaluate the techniques chosen for our study. Our user study aims to bridge some of the gap in understanding the circumstances that governs the decision making process in the presence of uncertain information.

3 UNCERTAINTY EVALUATION FRAMEWORK

Visualization techniques are usually designed to operate on data of a certain dimensionality. Sometimes techniques can be extended to operate on data of higher or lower dimensionality. For example, an isosurface is a 3D version of a 2D contour. We present an uncertainty evaluation framework that provides researchers with a structured classification to evaluate existing uncertainty visualization techniques. The framework was designed to consciously think of uncertainty from the perspective of the data being visualized and not by the uncertainty visualization technique employed. This framework also has the potential to provide a basis for developing better techniques and future user-studies (Fig 1).

Spatial data can be thought of as having zero, one, two, or three dimensions. In many applications, it is common to consider time as the outermost dimension. We decouple the temporal dimension and treat it specially because the temporal dimension usually has very different features and resolution than the spatial dimensions. It is worthwhile to note that this may not be true in the field of information visualization, where multi-dimensional data that is not spatial or temporal, is very common [40].
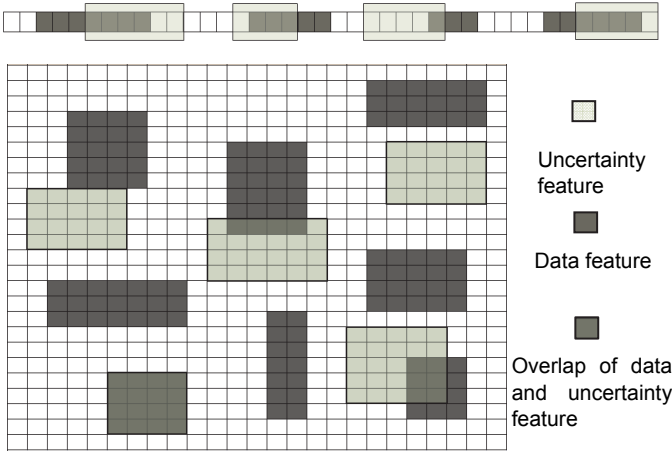
Fig. 2. Illustration of one of the layouts for data features and uncertainty features for 1D (top) and 2D (bottom) dataset generation. The datasets were designed to have some data features that overlapped with uncertainty features, and some that did not.

Fig. 3. Section of a 1D dataset depicting the generation of pseudo-observations. The green line is the true data A_{true}. The act of taking readings simulated by 50 normally distributed random numbers at each data value is shown by the randomly colored dots. One set of observations, A_k from the 50 observation sets A_0 to A_{49}, is illustrated by the red dotted line.

Further, scalars, vectors, and tensors can be thought of as three types of scientific visualization paradigms. Thus, data dimensionality (0D, 1D, 2D, 3D), visualization paradigm (scalar, vector, tensor) and the broad taxonomy of uncertainty visualization techniques (blurring, transparency, noise, etc.) form the three axes that define our classification (Fig 1). Our framework allows one to replace the technique axis with other classification schemes such as that of Pang et al. [31]. Additionally, the entire framework has a temporal axis. This is similar to the taxonomy proposed by Tory and Möller [38] but is more flexible, since this framework allows researchers to structurally extend the technique axis across other classification schemes and data dimensions.

Clearly, the problem space is huge. In this study, we considered a subset of the design space and focused on 1D and 2D data, considering four common uncertainty visualization techniques, namely scaling the size of glyphs, modifying the color attribute of glyphs, color-mapping the data surface with uncertainty, and adding errorbars, for scalar datasets.

4   STUDY DESIGN

This section presents the datasets used, the visualization techniques explored and the ones finally chosen, the participants, the pilot studies conducted, and various aspects of the main study. We borrowed design ideas from previous uncertainty visualization user studies [2, 10, 34, 41], as well as others, such as the 2D vector field visualization user-study by Laidlaw et al. [18], and the hurricane visualization user study by Martin et al. [27].

4.1   Data Generation

The data generation process was motivated by geoscience applications of visualization, which typically deal with various types of remotely sensed data, observed data at stations (e.g. buoys), data over a trajectory (e.g weather balloons), simulated weather data (e.g output from numerical models) and statistical studies (e.g. temporal correlations). Our objective was to design a controlled synthetic-data generation scheme that would be specific enough to provide immediate insight into geoscience uncertainty representation, as well as be generic enough to potentially have other applications.

We devised a mathematical method to simulate the data acquisition process and hence have complete control over the uncertainties introduced at different stages in a real data collection process. We simulated the process of repeated data collection, where, if any data measurement task is repeated a large number of times, the recorded values end up being normally distributed. If we take a subset of these values, we can derive a mean data value and a corresponding uncertainty value. We also introduced systematic uncertainty components that are an inherent part of any data collection process. Matlab [28] was used to generate the datasets. We first describe the process for the one-dimensional case and then extend it to two dimensions.
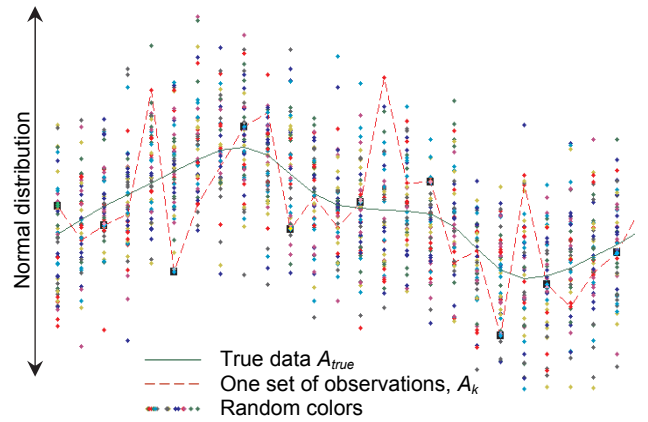
We begin with a 1D array, say A, consisting of 40 zeros and manually implant data features by setting consecutive index locations in A to a certain value representative of the signal strength, say S, at that location (dark regions in Fig 2). The value S was generated as a normally distributed random number (using the Matlab *randn* function) with a mean of 0.8 and a standard deviation of 0.2. Thus, we now have an array of zeros with user-defined data-features embedded in the array (Fig 2). Let this array be A'. In the next step, we interpolate the array A' (with the Matlab *interp1* function using cubic spline interpolation) to implant 3 points between every pair of array locations to generate 4 levels between them (Fig 2, 3). Let us call this array A_{true}. For our simulated dataset, A_{true} will represent the "truth value", which is analogous to the exact value of a continuous real-life phenomena such as the temperature of a place or water level of a sea surface that no instrument can ever record 'exactly'.

We then simulated the act of taking measurements or observations of the data. If a measurement is taken a large number of times, the errors in the observations can also be assumed to be normally distributed around the truth value assuming no systematic error, which constitutes the random uncertainty component. To simulate random uncertainty in our datasets, we generated 50 sets of readings, where every observation is normally distributed about values of the assumed true data A_{true} (Fig 3). Let these sets of values be A_0, A_1,..., A_{49}. To generate these, we first found the mean μ_{true} and standard deviation σ_{true} of the true data A_{true}. We used fractions and multiples of the standard deviation σ_{true}, say k, to generate three types of datasets having three ranges of random uncertainty in the data. Using the standard deviation of the truth data to generate multiple uncertainty levels seemed like a reasonable choice. We used three values for k which were 0.5, 1.0 and 1.5

For the dataset A_0, the ith element of A_0 corresponds to the ith element of A_{true} modified by the generated uncertainty..

Thus,

$$A_{0\,i} = A_{true\ i} + randn() * k * \sigma_{true}$$

We then took the first 10 observation sets, A_0, A_1,... , A_9, and calculated the mean and standard deviation for each index, generating a dataset say A''. This simulated the real-life step of averaging multiple data readings, and the standard deviation represented the uncertainty of the average.

An uncertainty study dataset is incomplete without a systematic uncertainty component. In real-life situations, often the uncertainty in the data exhibits patterns. This can be because of the nature of certain regions of the data, biases in the sensors, and a variety of other reasons. We introduced systematic biases in the generated random uncertainty by manually biasing certain sections of the array (Green regions in Fig 2). The bias values, say B', were normally distributed random values with a mean of 0.4 and standard deviation of 0.1 and were arbitrarily added or subtracted from the standard deviation values in the chosen sections of the array. This generated our uncertainty features.
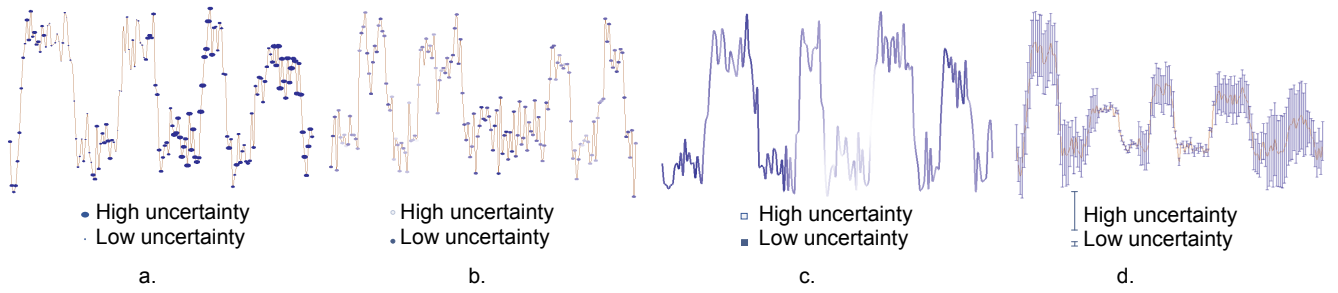
Fig. 4. Uncertainty visualization techniques explored for 1D datasets. a) Scaling the size of glyphs b) Altering the color attribute of glyphs c) Color-mapping the surface of data with uncertainty, and d) Using the traditional errorbars.
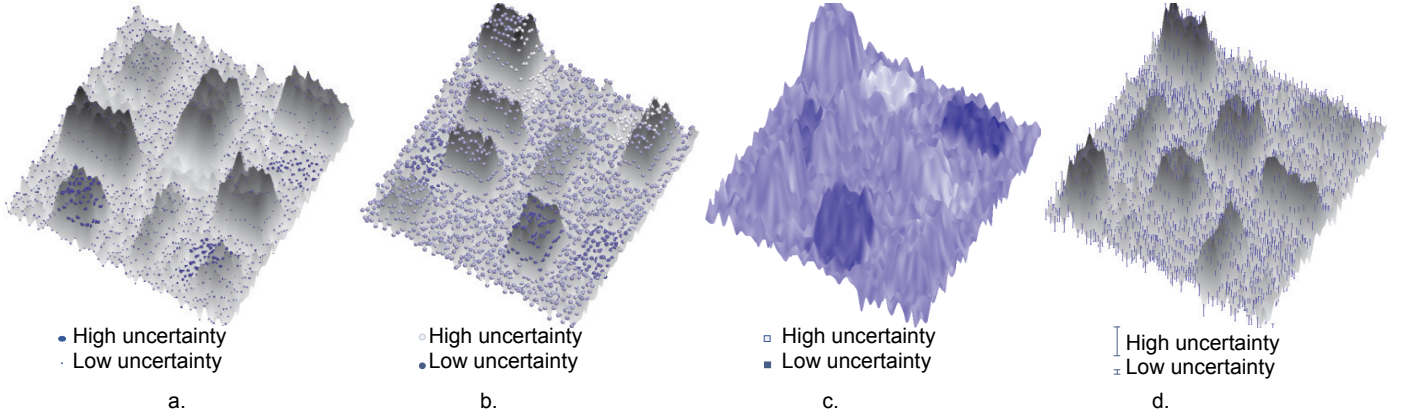

Fig. 5. Uncertainty visualization techniques explored for 2D datasets. a) Scaling the size of glyphs b) Altering the color attribute of glyphs c) Color-mapping the surface of data with uncertainty, and d) Using the traditional errorbars.

The 1D datasets had four different feature layouts (Fig 2), each with three values of k, making a total of twelve 1D datasets. We generated one extra dataset for use in the training of the participants.

The same logic was extended to generate 2D datasets (Fig 2). We started with a grid of 25×25 zeros and planted rectangular user-defined data features. The grid was interpolated (using Matlab *interp2* function) along both x and y axes to create 2 levels between every 2 grid points. 50 sets of pseudo-readings were generated in exactly the same way as that of the 1D case. We took the first 10 observation sets and generated the average signal value, the uncertainty value, and added rectangular uncertainty features. All parameters were kept the same as in the 1D case. Twelve 2D datasets were created for the main study and an additional dataset was created for the training module.

We do acknowledge that using real data from real sources has its merits, most notably being able to establish direct returns from the results of a user-study. We also acknowledge that not all data is normally distributed. We did not perform any tests on any real data or on other data distributions due to constraints of time. However, in the future, we intend to conduct a similar experiment on such data and evaluate how the results correlate.

4.2 Visualization Techniques

Using our uncertainty visualization framework, we chose four visualization techniques that could be applied to both 1D and 2D data. These were scaled sizes of glyphs, altering the color attribute of glyphs keeping the size constant, color-mapping the data surface with the uncertainty, and traditional errorbars (Fig 4, 5). The data was displayed in greyscale except where color-mapping of the surface was used. The 2D data surface was rendered with an orthographic projection to minimize 3D perspective effects that may interfere with perception of height. This ensured that the uncertainty representations would be of uniform size regardless of the distance from the eye. A few other techniques such as smooth and striped gradients, animation of glyphs, and animation of the data surface were explored but were not included in the user study (Fig 6).

There were two considerations in removing some of our suggested techniques from the final study, the first being the inherent merit of the technique and the second being the number of questions it would add to the study. Smooth and striped gradients were the first to be eliminated because they display incorrect uncertainty information across steep slopes. In the 1D case (Fig 6a, b), the uncertainty information in the gradient upon a steep slope aligns itself with the slope, resulting in a thin ribbon. Forcing the ribbon to be always orthogonal is not an elegant solution.

In the 2D case, the surface animation would either hide the data surface, or be itself hidden by the data surface. We eliminated animation of glyphs also because of similar occlusion issues in the animation. To be consistent, when we had to eliminated a technique, we eliminated it for both 1D and 2D datasets. We did not want to test a technique that worked for say 1D and did not for the 2D case.

The second concern for us was user fatigue. We did not want to overwhelm the user with too many questions of the same type, which could jeopardise the quality of our results.

The display area had a size of 800×800 pixels. The interpretation of visualizations with the errorbars was straightforward [7]. Small bars (smallest being about 8 pixels tall in 1D and about 3 pixels tall in 2D) represented low uncertainty while large bars (largest being about 85 pixels tall in 1D and about 40 pixels tall in 2D) represented high uncertainty. When scaled glyphs were used, large glyphs (largest being about 10×10 pixels in both 1D and 2D) represented high uncertainty and small glyphs (smallest being about 3×3 pixels in both 1D and 2D) represented low uncertainty. We used flat shading on the glyphs for the 1D dataset, however, we enabled lighting for the 2D dataset to give users a sense of location of the glyphs. For altering the color attribute of glyphs as well as color-mapping of the surface, we mapped a low uncertainty value to a saturated shade of blue and mapped a high uncertainty value to an unsaturated shade of blue. This is also the scheme suggested by MacEachren [25] and Hengl [11]. The mapping is opposite to the intuitive notion of high and low; however, here we are dealing with negatives, for example 'high uncertainty' implies low certainty.

Shades of blue were chosen to convey the uncertainty since blue is a cool color and would cause minimum visual fatigue over the duration of the experiment. Red was used as a preattentive cue to   mark regions of interest

High uncertainty: wide ribbon
Low uncertainty: thin ribbon

a.

High uncertainty: wide gradient
Low uncertainty: thin gradient

b.

High uncertainty: large amplitude
Low uncertainty: small amplitude

c.

High uncertainty: large amplitude
Low uncertainty: small amplitude

d.

High uncertainty: tall bars
Low uncertainty: short bars

e.

High uncertainty: large amplitude
Low uncertainty: small amplitude

f.

High uncertainty: large amplitude
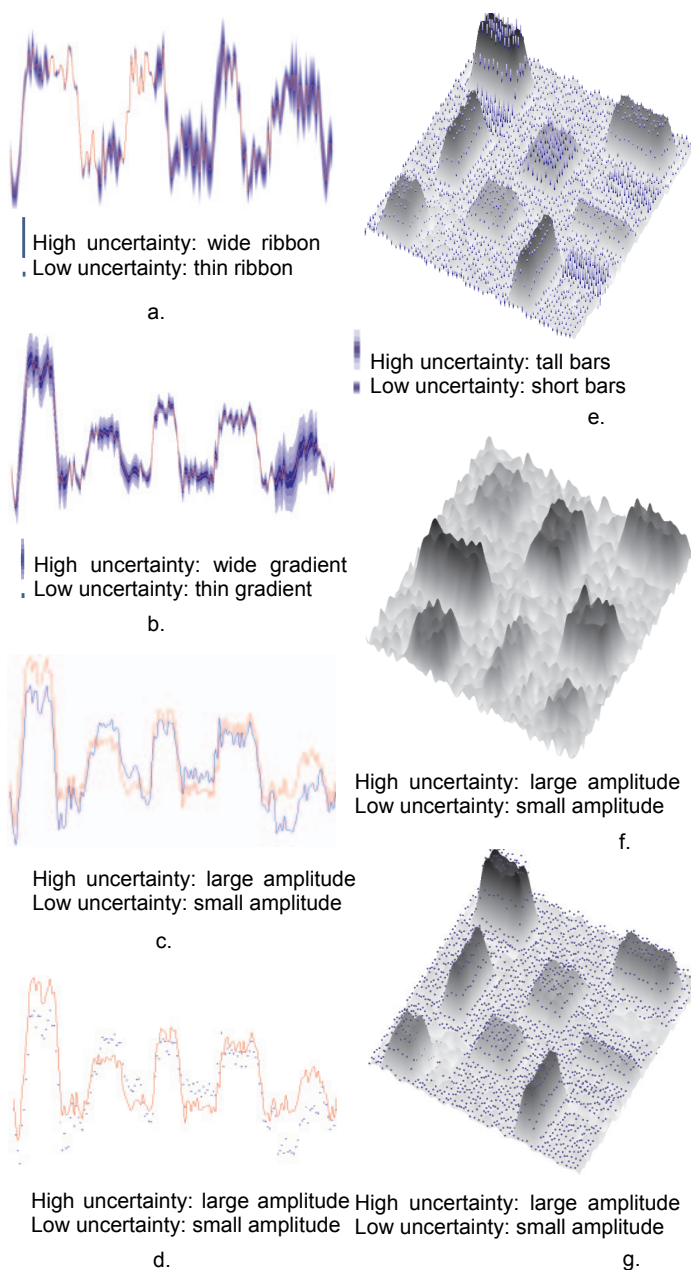Low uncertainty: small amplitude

g.

Fig. 6. Other uncertainty visualizations explored but not used in our experiments. For 1D: a) Applying a color gradient to the uncertainty range b) Applying a striped gradient c) Animation of a line in the uncertainty range d) Animation of glyphs. For 2D: e) Using the striped bars f) Animation of the surface, and g) Animation of glyphs.

and highlight user selections [40]. A legend was always provided to aid the user.

4.3 Participant Pool

The participants of our user study were mostly graduates and under-graduates of Mississippi State University. We also had two senior participants who are researchers at the university. We had a total of 36 participants, of which 3 participated in a trial run, 6 participated in a pilot study, and the remaining 27 participated in the main study. Of the 36 participants, 27 were male and 9 were female. None of the participants reported color-blindness while 17 reported 20/20 corrected vision. Most of the participants had some understanding of statistics and used charts and graphs for their day-to-day

activities although none of these skills were set as prerequisites to participating in the study. Most users typically spent more than 15 hours weekly using a computer. Each participant was paid $10 for their time and participation.

4.4 User-study tasks

Not much is known about how domain scientists perceive and use uncertainty. We consulted Dr. Jamie Dyer, a meteorologist, to determine what might be a real world scenario where uncertainty would be a part of his decision making process. He had temperature data in mind and indicated that he would be interested in looking at regions of extreme (high or low) uncertainty. He also wanted to be able to discern features in the data, in the presence of uncertainty. Keeping this in mind, we designed two types of tasks: searching tasks and counting tasks. The searching tasks primarily explored the perception of random uncertainty while the counting tasks explored the perception of systematic uncertainty, along with the cognizance of the underlying data. Dr. Dyer mentioned that he liked to look at the entire data and then focus on a region of interest. The searching and counting tasks were designed to simulate such an exploratory navigation of the data.

4.4.1 Search tasks

The search tasks involved searching for locations of high or low uncertainty from within an area marked in red (Fig 7). The entire dataset was always shown to the user. This was done keeping real-life data exploration tasks in mind. Any spot within the marked region could be selected by the user and the corresponding data/uncertainty values would be interpolated whenever necessary. This design decision was made keeping data collection in geosciences in mind, where we take samples at specific locations over a domain and then interpolate if we need values in between.

We expected users to perform similarly in both the search tasks, however, our results indicate that there was a significant difference as discussed in our results section. These tasks had more than one correct answer. A user response was considered correct if the chosen location had an uncertainty value within the top 10th percentile of the entire range of uncertainty for a task requiring the user to find the location of highest uncertainty, or the bottom 10th percentile for a task requiring the user to find the location of lowest uncertainty. We had also tested with the 5th percentile but felt that it made the tasks too difficult to perform reasonably. The 10th percentile seemed to be a reasonable balance between making the user study impossibly difficult and too easy. Although we did not perform a formal test, we expect the results to remain the same empirically.

Location and proximity of high and low uncertainty areas had an effect on the correctness of the user responses. If a region of interest included both a high and a low uncertainty feature, the range of uncertainty values was much larger than had there been just one uncertainty feature or no uncertainty feature. As a result the number of correct answers changed on a case by case basis. Locating a spot of high or low uncertainty was thereby facilitated, however, a correct answer was not guaranteed by choosing just any location within a feature. There were variations within the feature too, and a user had to make an informed decision as opposed to a blind selection within an approximate region. To control arbitrary effects, we designed the regions of interest to uniformly include high, low, both and none of the uncertainty features.

4.4.2 Counting tasks

The counting tasks involved counting either data features or counting uncertainty features within an area marked in red (Fig 8). The definition of a data-feature in our study is the presence of any 'peak' in the data. Artifacts resulting from the introduction of systematic uncertainty were called uncertainty features in this study, which manifest as regions of extreme glyph-size, glyph-color, errorbar size or surface-color.

One might argue against the merit of having a counting task for data features in an uncertainty visualization experiment. We contend that it is generally important for a user to be always aware of the data and the counting tasks would evaluate the effectiveness of the techniques in retaining a sense of the data.

4.5 Interface Design

For the search questions, the interface provided one slider for the 1D data and two sliders for the 2D data to navigate a small red highlight to the chosen answer location (Fig 7). Clicking the sliders displayed cross-hair guides to ease the navigation. We were time constrained to implement direct object-picking in our interface. We eventually found that users were very comfortable using this interaction metaphor and could reach the desired screen location with at most 2-3 movements of the sliders.

For the counting questions, radio buttons with four static answer choices of 0, 1, 2 and 3 were provided (Fig 8). In these questions, the sliders were hidden and four radio button options placed vertically were displayed. Other layouts for the radio buttons were not experimented. Users were expected to make a selection from one of the four radio-button choices. The regions of interest were designed in such a way as to never exceed three data or uncertainty features, and were uniformly designed to include all possibilities. We feel that having a fixed set of choices makes the quality of responses better than having users enter a numeric digit on a prompt. This also ensures a consistent response structure across the methods.

Users could not skip a question. They clicked an 'Accept' button to record their answer and response time, and only then could they click a 'Next' button to go to the next question. A break of 5 minutes was given after every 30 questions.

A trial run identified weaknesses with our initial design. Of the three participants in the trial run, two had prior experience designing user studies and their debriefing proved helpful in improving the design. Most notably, the rotation of the 2D surface along the z-axis was fixed to 30 degrees from the original 45 degrees to alleviate some of the artifacts resulting from the view-aligned overlap of errorbars and glyphs. Additionally, the range of the sliders was adjusted to restrict the navigation of the highlight to within the marked region. Users were not allowed to rotate the view or zoom in.

4.6 Participant Training

We typically spent about 15 to 20 minutes to brief the participant about the user study. This involved getting their informed consent and completing a general questionnaire, followed by an explanation of the tasks expected of them. Users were then assigned a computer which ran a training module which was a variation of the software used in the actual study. It familiarized users with the interface and posed 8 questions, one of each type, on the two training datasets. The software highlighted the correct answers to the users to give them feedback on their performance. No person was involved in this process. We felt that users were confident to take on the main user-study after this exercise.
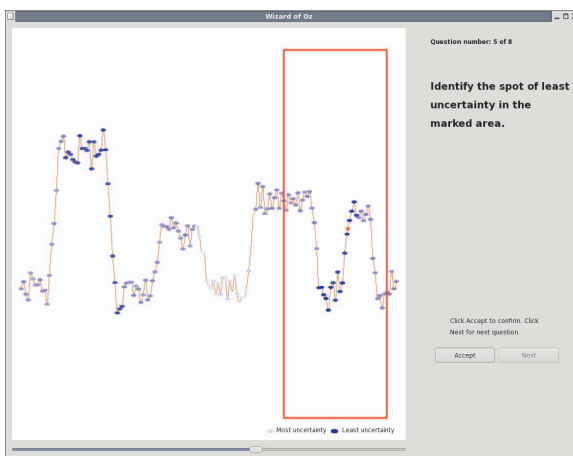
4.7 Identifying free parameters

A pilot study was conducted to identify the free parameters in our user-study. These were the size of errorbars and the size of the glyphs. We had 6 participants but we could use data from only 4 of the participants. The quality of the answers from the other 2 was unacceptable because, primarily, they seemed unmotivated and finished too soon. Also, the correctness of their responses was about 50% lower than the others. For this pilot-study, we used three sizes of errorbars and glyphs to compare small, medium, and large representations. The largest of the glyphs was limited to not exceed the size of the grid-cell.

Each participant was asked 144 questions in random order. Although it is difficult to draw meaningful inferences from data from just 4 users, we did find trends that helped us make reasonable assumptions. Users found it easiest to use the smallest errorbars in both 1D and 2D and so we chose to use errorbars of the smallest dimensions in the main study. For glyph color-mapping of the 1D data, users found it easiest to use glyphs that had the largest size among the three evaluated sizes. For the 2D data, the three glyph sizes used for glyph color-mapping did not show such a clear trend but had the minimum variance in accuracy of responses for the largest size. So we chose the largest size of the glyphs for use in the main study.

Unlike color-mapping, we did not observe any trends in the responses for the small, medium, and large size ranges used with the glyph-size technique. We attribute it to the inherent nature of the mapping of uncertainty values to size of glyphs making it difficult to find a separation between the three maximum sizes. So, for this technique, we resorted to using the size we chose for glyph color-mapping.

4.8 The Main Study

We had 27 participants in the main study, each answering 96 questions of which 48 questions were on the 1D datasets and the remaining 48 on the 2D datasets. Each set of 48 questions, consisted of three sets of 16 questions each based on data generated using one of the three k values (0.5, 1.0 and 1.5). The 16 questions asked formed a complete 4×4 design of the four visualization techniques explored and the four user-tasks chosen. The response time in milliseconds was recorded for each question. Each user was presented a different shuffled order of questions. The four questions asked were:

- How many data features are present in the marked area?
- How many uncertainty features are present in the marked area?
- Identify the spot of least uncertainty in the marked area.
- Identify the spot of most uncertainty in the marked area.

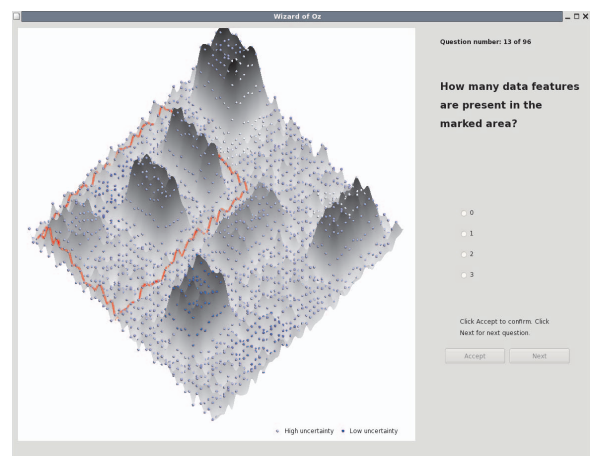Fig. 7. User interface for a search task on a 1D dataset.

Fig. 8. User interface for a counting task on a 2D dataset.

5   METHOD OF ANALYSIS

Every correct answer was given a score of 1 and every incorrect answer was given a score of 0. Since there were three questions for a given visualization technique per user task, a participant could achieve a maximum score of 3 for the task, given a visualization technique. Score summaries were created separately for the 1D and 2D datasets.

For each dataset, a 4×4 full factorial ANOVA was computed to assess the differences in performances for different questions and different techniques [29]. The summary ANOVA table indicated a significant interaction between type of tasks and techniques used (F (9, 416) = 9.968, $p < .0001$) for the 1D dataset. For the 2D dataset, the summary ANOVA table also indicated a significant interaction between type of tasks and techniques used (F (9, 416) = 7.818, $p < .0001$). This implied that whether there was a significant difference between techniques or not, depended on the type of tasks assigned to the subjects, for both the datasets. Thus, to further explore the results, 8 one-way ANOVAs were computed to capture the Simple Main Effects for each dataset.

The first 4 one-way ANOVAs were intended to see if there were any statistically significant differences between the 4 techniques used with respect to the user tasks. All possible pairwise comparisons were made (6 pairwise comparisons) between the techniques to see if any technique was visibly superior to the rest by creating contrast coefficients to test for the significance of each comparison. The alpha level was set at 0.0083 as compared to widely accepted 0.05 after using Bonferroni's correction (α/c; where c = number of comparisons) to control for Type I error. Also, the t-test value which does not assume equality of variances was reported for each comparison since the data indicated slight violation of homogeneity of variances. This was computed for the 4 task types, viz. Search Low Uncertainty Locations (SLU), Search High Uncertainty Locations (SHU), Count Uncertainty Features (CUF) and Count Data Features (CDF). The specific findings are listed in Table 1. We only report the statistically significant results.

The next 4 ANOVAs were intended to see if there were any statistically significant differences between the 4 user tasks with respect to the visualization techniques used, viz, Glyph-size, Glyph-color, Surface-color and Errorbars. The specific findings are listed in Table 2.

We ran our core statistical methods (4×4 full factorial ANOVA) on the obtained scores, from which we identified significantly better performing techniques and tasks. Similar ANOVAs could be based on the analysis of time, however, that would have inordinately complicated the reporting in the time and space available to us which is why we chose graphical techniques to illustrate the time performance (Figs 9b, 10b, and 11b). Also, we were more interested in the accuracy assessment than the time performance in our research goal.

6   RESULTS AND DISCUSSION

We found a consistent trend in the accuracy of responses and the response time for questions on the 1D and 2D datasets (Fig 9, 10, 11). Since we found a statistically significant interaction between the techniques used and the user-tasks, inferences about a general order of performance of the techniques could not be found, however, we did find several interesting discoveries that we think are useful for uncertainty visualization design. In both cases, Errorbars performed significantly worse than the other techniques studied even though it took substantially more time to answer these questions. One possible explanation of the difference between the errorbars and glyphs is the difference in area between the two representations. We had tested with different errorbar sizes and glyph sizes only in our pilot study from which we determined the most effective size to use in the main study. It would be interesting to test with other shapes of glyph as well.

Visualization researchers agree that the choice of a visualization technique is heavily context dependent. All the visualizations in the study have the same data-density. So, they are fair in the sense that all the techniques were being compared vis-à-vis the same conditions. It is also possible that the data density played a role in perception leading to the poor performance of errorbars. This may be taken as a valuable lesson in designing visualizations for both 1D and 2D cases, which have data-

Table 1. ANOVA results for pairwise comparison on techniques

Data Dim	Task t- values (df)	Significantly better	Significantly worse
1D	Search for location of low uncertainty (SLU) F (3, 104) = 16.176, $p < .0001$	Glyph size**	Glyph color
		Glyph size**	Surface color
		Glyph size*	Error bars
		Error bars*	Surface color
	Search for location of high uncertainty (SHU) F (3, 104) = 13.874, $p < .0001$	Glyph color**	Glyph size
		Glyph color**	Error bars
		Surface color**	Glyph size
		Surface color**	Errorbars
	Count uncertainty features (CUF)	No significant difference	
	Count data features (CDF)	No significant difference	
2D	Search for location of low uncertainty (SLU) F (3, 104) = 6.775, $p < .0001$	Surface color*	Glyph size
		Surface color*	Glyph color
		Surface color*	Error bars
	Search for location of high uncertainty (SHU) F (3, 104) = 48.144, $p < .0001$	Surface color*	Glyph size
		Surface color**	Error bars
		Glyph size**	Error bars
		Glyph color**	Error bars
	Count uncertainty features (CUF) F (3, 104) = 7.534, $p < .0001$	Surface color*	Glyph size
		Surface color*	Glyph color
		Surface color*	Error bars
	Count data features (CDF) F (3, 104) = 25.910, $p < .0001$	Glyph size*	Glyph color
		Glyph size**	Surface color
		Glyph size*	Error bars
		Glyph color**	Surface color
		Error bars**	Surface color

* p ≤ .0083
** p ≤ .0001

Table 2. ANOVA results for pairwise comparison on tasks

| Data Dim	Technique t- values (df)	Significantly better	Significantly worse
1D	Glyph size F (3, 104) = 13.499, $p < .0001$	SLU**	SHU
		SLU**	CUF
		CDF**	SHU
	Glyph color F(3, 104) = 6.880, p < .0001	SHU*	SLU
		CDF**	SLU
		CDF*	CUF
	Surface color F (3, 104) = 17.295, $p < .0001$	SHU**	SLU
		CDF**	SLU
		CDF*	CUF
	Error bars F (3, 104) = 11.587, $p < .0001$	CDF**	SHU
		CDF**	CUF
2D	Glyph size F(3, 104) = 16.721, p<.0001	SHU**	SLU
		SLU**	CUF
		CDF*	SHU
		CDF**	SLU
		CDF**	CUF
	Glyph color F (3, 104) = 11.780, $p < .0001$	SHU**	SLU
		SLU*	CUF
		CDF**	SLU
	Surface color F(3, 104) = 36.356, $p < .0001$	SHU*	SLU
		SHU**	CDF
		CUF*	CDF
		CUF*	SLU
		SLU**	CDF
	Error bars F(3, 104) = 14.067, $p < .0001$	SLU**	SHU
		CUF**	SHU
		CDF**	SHU

* p ≤ .0083
** p ≤ .0001
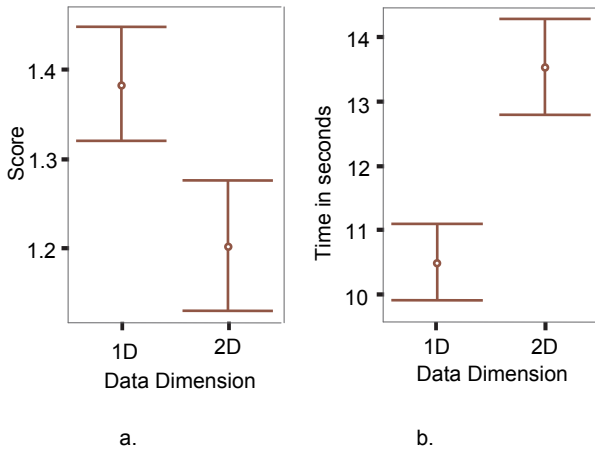
Fig. 9. Overall performance of the two datasets. a) Average scores attained for each dataset, and b) Response time recorded for each dataset.

Fig. 10. Performance plots for the 1D datatset. a) Average scores for each technique, and b) Response times recorded for each technique.

Fig. 11. Performance plots for the 2D datatset. a) Average scores for each technique, and b) Response times recorded for each technique.

densities comparable to our datasets. However, we do not have sufficient anecdotal evidence that might help us understand this.

The first sets of pairwise comparisons were between the different uncertainty visualization techniques for different user tasks.

For the 1D tasks (Table 1), users performed significantly well using Glyph-size when the task was to search for locations of least uncertainty. However, both Glyph-color and Surface-color performed better than Glyph-size when the task was to search for locations of high uncertainty. We did not expect to find a significant difference between the two search tasks since both the tasks were designed to find extremes in the data. This leads us to believe that human perception of uncertainty ranges using Glyph-size, Glyph-color and Surface-color may not be uniform. Our mapping between visual features and uncertainty was linear. One explanation is that the uncertainty was not translated linearly to the visual features and hence the difference in performance between the two search tasks.

For the 2D tasks (Table 1), Surface-coloring performed reasonably well for all questions except counting of uncertainty features. Since shape of the surface was the primary visual cue for data features, we feel that color-mapping of the data surface with uncertainty reduces some of the strength of the shape information. Hence, we see that all other techniques outperform Surface-color for the counting of data features task. On the whole it might sound encouraging to use Surface-coloring to represent uncertainty. While this may work well, one must be aware that it reduces a user's awareness of the actual data. Glyph-size and Glyph-color performed somewhat better than Error-bars although they were worse than Surface-color which bolsters the argument for using one of them as a reasonable trade-off.

The second sets of comparisons were between the tasks for different uncertainty visualization techniques (Table 2).

For the 1D techniques, Searching for Low Uncertainty (SLU) was clearly the easiest task to perform when the technique was Glyph-size. However, Searching for High Uncertainty (SHU) task was significantly easier when the technique was Glyph-color. Interestingly, for the 2D data, Searching for High Uncertainty (SHU) was consistently easier than Searching for Low Uncertainty (SLU) for all techniques except for Errorbars, where Searching for Low Uncertainty (SLU) was significantly easier.

We found that Counting Data Features (CDF) was generally more accurate except for the 2D case when surface-coloring was used.

On the whole, we found that it took consistently longer for users to respond to questions on the 2D datasets than to questions on the 1D datasets. The accuracy of responses was also higher for the 1D dataset (Fig 9). This is not surprising because 2D tasks are generally more difficult than 1D tasks.

Our uncertainty visualization user study brings to light several interesting observations. One such result is that user efficiency in the two search tasks that are opposite of one another (locations of high uncertainty vs. locations of low uncertainty) are significantly different. This is contrary to common
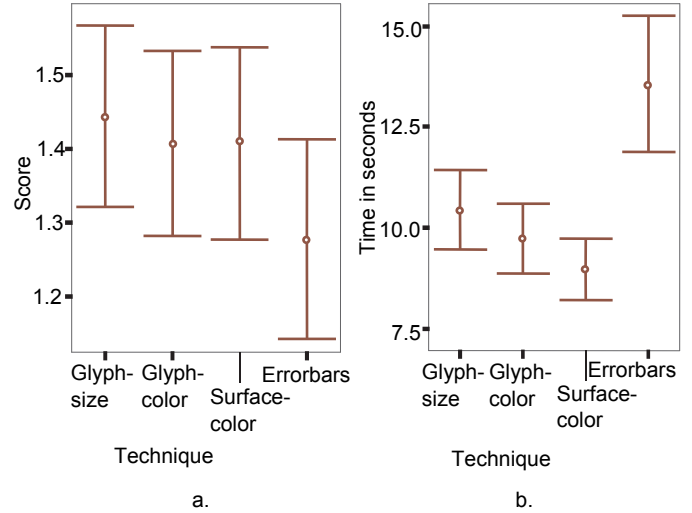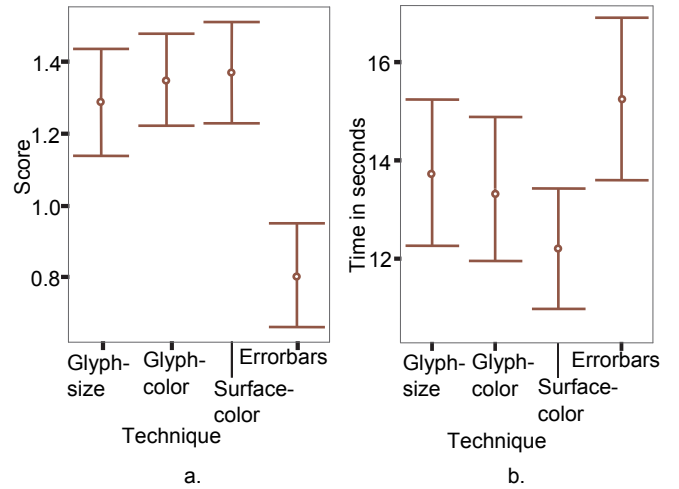
understanding and may be attributed to a non-linear perception of the mapping between uncertainty and the visual metaphor. This may drive us to find techniques that compensate for our perceptive biases, or design techniques that are unbiased. One such technique is the Linearized Optimal Color Scale introduced by Levkowitz and Herman [20].

Another aspect that merits discussion is the cognitive associability of high uncertainty with faint colors. The term high uncertainty also associates well with large glyph-sizes. We experimented by reframing our questions and the legend with terms like 'high certainty' and 'least uncertainty', and eventually stuck with 'high uncertainty' and 'low uncertainty' as it seemed to facilitate the cognitive mapping.

We feel that some of our results should force us to think again about the techniques we use on a daily basis. Our study questions the effectiveness of the almost universally used errorbars in data visualization. Alternative methods may be suitable in many cases.

7   CONCLUSION

In this paper, we presented a user study to compare effectiveness of four uncertainty visualization techniques on 1D and 2D datasets. We had hoped that we would be able to provide guidelines for information design with uncertainty from the results of our study, but our results are such that we

cannot identify clear winners. We did not find a consistent order among the four techniques for all the tasks, although the particular findings could be useful for uncertainty visualization design. Errorbars consistently performed poorly compared to the other evaluated techniques. The accuracy of responses for 1D tasks was higher than that of the 2D tasks although 2D tasks consistently took longer to respond to. We also found that effectiveness of uncertainty visualization techniques were highly dependent on the task at hand. User efficiency between the two search tasks was significantly different from one another which raised interesting questions. Surface-coloring worked well except for counting 2D data features. Performance of Glyph-size and Glyph-color seemed reasonable. We feel our results could help researchers in their choice of uncertainty visualization technique for their scientific tasks.

We also presented a method to create synthetic data with uncertainty. Additionally, we presented an Uncertainty Evaluation Framework which provides a structured design environment using which researchers can create effective uncertainty visualizations across various visualization paradigms. We expect our findings to be useful for researchers who have a need of displaying dense 1D or 2D data with uncertainty. In particular, we see applications from geoscience such as visualization of severe weather outbreaks, precise terrain modeling, and moving front locations to benefit from this study.

8 FUTURE WORK

Our results on differences between uncertainty visualization methods motivate future research in this area. With our Uncertainty Evaluation Framework, we plan to use the results from this study to guide our future uncertainty visualization endeavours. Perceptual research to identify the reason why the two search tasks differed so significantly could be potentially beneficial. It may also to be very useful to research how experts use uncertainty in their decision making process and design experiments around such observations. We also plan to extend this research to evaluate uncertainty visualization techniques for 3D data as well as time series data. This may prove beneficial for users of data that are inherently 3D and have samples in time. Weather researchers for example may significantly benefit from such knowledge.

9 ACKNOWLEDGMENTS

REFERENCES

[1] J. Bertin. Semiology of graphics: Diagrams, networks, maps. (Translation from a French 1967 edition by W. Berg, editor). Madison, WI: University of Wisconsin Press. 1983.

[2] S. Blenkinsop, P. Fisher, L. Bastin, and J.Wood. Evaluating the Perception of Uncertainty in Alternative Visualization Strategies. *Cartographica* 37(1):1-13. 2000.

[3] B. P. Buttenfield, Representing data quality. *Cartographica* (special content, Mapping Data Quality) 30(2&3):1-7. 1993.

[4] A. Cedilnik and P. Rheingans. Procedural annotation of uncertain information. In *Proc., Visualization 2000,* Salt Lake City, 77-84, 2000.

[5] CIPM, BIPM. In *Proc., Verb. Com. Int. Poids et Mesures* (in French). 1982.

[6] H. Couclelis. The certainty of uncertainty: GIS and the limits of geographic knowledge. In *Trans. in GIS* 7(2): 165-75, 2003.

[7] G. Cumming, F. Fidler, and D. L. Vaux. Error bars in experimental biology. *Journal of Cell Biology*, 177:7-11. 2007.

[8] T. J. Davis and C. P. Keller. Modeling and visualizing multiple spatial uncertainties. *Computers & Geosciences* 23(4): 397-408. 1997.

[9] N. Gershon. Short note: Visualization of an imperfect world. In *IEEE Computer Graphics and Applications*, 18: 43–45, Jul.-Aug. 1998.

[10] M. Harrower. Representing uncertainty: Does it help people make better decisions? In *UCGIS Workshop: Geospatial Visualization and Knowledge Discovery Workshop*, National Conference Center, Landsdowne, VA., 2003, (invited white paper).

[11] T. Hengl. Visualisation of uncertainty using the HIS colour model: Computations with colours. In *Proc., 7th International Conference on GeoComputation*, Southampton, United Kingdom. 8-17. 2003.

[12] T. Hengl and N. Toomanian. Maps are not what they seem: representing uncertainty in soil property maps. In *Proc., 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, 805-813. 2006.

[13] S. Huang. Exploratory Visualization of Data with Variable Quality. Masters Thesis, Worcester Polytechnic Institute, Jan. 2005.

[14] G. J. Hunter and M. F. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. In *Proc., URISA*, Atlanta, July 25–29. pages 1–14. 1993.

[15] B. Jiang, F. Ormeling, and W. Kainz. Visualization support for fuzzy spatial analysis. In *Proc,. ACSM/ASPRS Conference*, Charlotte, North Carolina. 291-300, 1995.

[16] C. R. Johnson ad A. R. Sanderson. A Next Step: Visualizing Errors and Uncertainty. *Visualization Viewpoints*, Sep.-Oct. 2003.

[17] D. Kobus, S. Proctor, and S. Holste. Effects of experience and uncertainty during dynamic decision making. *International Journal of Industrial Ergonomics.* 28:275-90. 2001.

[18] D. H. Laidlaw, R. M. Kirby, C. D. Jackson, J. S. Davidson, T. S. Miller, M. da Silva, W. H. Warren, and M. J. Tarr. In *IEEE Trans. Visualization and Computer Graphics*, 11(1), Jan/Feb 2005.

[19] M. Leitner and B. P. Buttenfield. Guidelines for the display of attribute certainty. *Cartography and Geographic Information Science.* 27(1):3–14. 2000.

[20] H. Levkowitz and G. T. Herman. Color scales for image data. In *IEEE Computer Graphics and Applications*, 12(1), pages 72-80, Jan 1992.

[21] H. Li, C. W. Fu, Y. Li, and A. Hanson. Visualizing Large-Scale Uncertainty in Astrophysical Data. In *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1640-1647, Nov.-Dec. 2007.

[22] S. K. Lodha, C. M. Wilson, and R. E. Sheehan. LISTEN: Sounding uncertainty visualization. In *Proc., IEEE Visualization '96*, San Francisco, California. 189-95. 1996.

[23] K. Lowell. Outside-in, inside-out: Two methods of generating spatial certainty maps. In *Proc., Second Annual Conference of GeoComputation*, University of Otago, Dunedin, New Zealand, 15-25. 1997.

[24] C. Lundstrom, P. Ljung, A. Persson, and A. Ynnerman. Uncertainty Visualization in Medical Volume Rendering Using Probabilistic Animation. In *IEEE Trans. Visualization and Computer Graphics*, 13(6):1648-1655, Nov.-Dec. 2007.

[25] A. M. MacEachren. Visualizing Uncertain Information. *Cartographic Perspective*, 13:10-19, 1992.

[26] A. M. MacEachren, C. A. Brewer, and L. W. Pickle. Visualizing georeferenced data: Representing reliability of health statistics. *Environment and Planning*: A30:1547-61. 1998.

[27] J. P. Martin, J. E. Swan II, R. J. Moorhead II, Z. Liu, and S. Cai. Results of a User Study on 2D Hurricane Visualization. *Computer Graphics Forum (Special Issue on EuroGraphics / IEEE VGTC EuroVis'08, formerly VisSym).* 27(3): 991-998. 2008.

[28] Matlab, The Mathworks Inc., R2008a, Natick, Massachusetts, USA.

[29] S. E. Maxwell and H. D. Delaney. Designing experiments and analyzing data: A model comparison perspective. *Lawrence Erlbaum Associates Inc. Publishers*, New Jersey. 2004.

[30] C. Olston and J. D. Mackinlay. Visualizing Data with Bounded Uncertainty. In *Proc., IEEE Symposium on Information Visualization 2002 (InfoVis'02).* 37-40. 2002.

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

[32] P. Rheingans and S. Joshi. Visualization of Molecules with Positional Uncertainty. In *Proc., Data Visualization '99.* 299-306. 1999.

[33] G. S. Schmidt, S. L. Chen, A. N. Bryden, M. A. Livingston, L. J. Rosenblum, and B. R. Osborn. Multidimensional Visual Representations for Underwater Environmental Uncertainty. *IEEE Computer Graphics and Applications* 24(5): 56-65. 2004.

[34] D. M. Schweizer and M. F. Goodchild. Data quality and choropleth maps: An experiment with the use of color. In *Proc., GIS/LIS '92*, San Jose, California. ACSM and ASPRS, Washington, D.C., 686-99. 1992.

[35] T. Strothotte, M. Masuch, and T. Isenberg. Visualizing Knowledge about Virtual Reconstructions of Ancient Architecture. In *Proc., International Conference on Computer Graphics*, IEEE Computer Society. 36–43. 1999.

[36] B. N. Taylor and C. E. Kuya. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. NIST Technical Note 1297, Physics Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, 1994.

[37] J. Thomson, B. Hetzler, A. MacEachren, M. Gahegan, and Misha Pavel. A Typology for Visualizing Uncertainty. In *Proc., Conference on Visualization and Data Analysis 2005* (part of the *IS&T/SPIE Symposium on Electronic Imaging 2005*). San Jose, California. Jan. pages 16-20. 2005.

[38] M. Tory and T. Möller. Rethinking Visualization: a High-Level Taxonomy. In *Proc., IEEE Symposium on Information Visualization 2004 (Infovis '04)*. 151-158. 2004.

[39] E. R. Tufte. The Visual Display of Quantitative Information. Graphics Press. 1986.

[40] C. Ware. Information Visualization: Perception for Design. Morgan Kaufmann Publishers, 2000.

[41] S. P. Wechsler. Perceptions of Digital Elevation Model Uncertainty by DEM Users. *URISA Journal*, 15:57-64. 2003.

[42] T. Zuk and M. S. T. Carpendale. Theoretical Analysis of Uncertainty Visualizations. In *Proc., SPIE & IS&T Conf. Electronic Imaging*, Vol. 6060: Visualization and Data Analysis 2006, 606007. 2006.