Visualizing Data with Bounded Uncertainty

Chris Olston*
Stanford University
olston@cs.stanford.edu

Jock D. Mackinlay
Palo Alto Research Center
mackinlay@parc.com

Abstract

Visualization is a powerful way to facilitate data analysis, but it is crucial that visualization systems explicitly convey the presence, nature, and degree of uncertainty to users. Otherwise, there is a danger that data will be falsely interpreted, potentially leading to inaccurate conclusions. A common method for denoting uncertainty is to use error bars or similar techniques designed to convey the degree of statistical uncertainty. While uncertainty can often be modeled statistically, a second form of uncertainty, bounded uncertainty, can also arise that has very different properties than statistical uncertainty. Error bars should not be used for bounded uncertainty because they do not convey the correct properties, so a different technique should be used instead. In this paper we describe a technique for conveying bounded uncertainty in visualizations and show how it can be applied systematically to common displays of abstract charts and graphs. Interestingly, it is not always possible to show the exact degree of uncertainty, and in some cases it can only be displayed approximately.

Keywords: uncertainty visualization, bounded uncertainty

1 Introduction

In most data-intensive applications, uncertainty is a fact of life. For example, in scientific applications, error-prone measurements or incomplete sampling often result in uncertain data. Another example is financial analysis, where it is common for some data to represent uncertain projections about future behavior. Even when it is possible to gather precise data, there are many real-time applications, such as network monitoring, mobile object tracking, and wireless ecosystem monitoring, in which uncertainty may be introduced intentionally to conserve system resources while data is being transmitted or processed. When data is uncertain, it is critically important that analysis tools, including information visualization tools, make users aware of the presence, nature, and degree of uncertainty in the data as these factors can greatly impact decision-making. If users are misinformed about the uncertainty associated with their data, they may draw inaccurate conclusions, potentially leading to costly mistakes.

A report by the US Department of Commerce National Institute of Standards and Technology (NIST) [TK94] identifies two predominant forms of uncertainty, which we call *statistical uncertainty* and *bounded uncertainty*. Statistical and bounded uncertainty have dramatically different meanings. Statistical uncertainty is typically captured by a potentially infinite distribution of possible values with a peak indicating the most likely estimate. In contrast, with bounded uncertainty no distribution of values can be assumed, but the exact value is known to lie inside an interval defined by precise lower and upper bounds.

Pang et al. [PWL97] argue, as we do, that uncertainty should be presented along with data in visualization applications. After discussing traditional techniques for showing statistical uncertainty such as error bars, they propose an extensive suite of techniques for conveying uncertainty in scientific visualization applications. Many of these techniques can be adapted to information visualization scenarios. However, techniques for conveying statistical uncertainty tend to be misleading when used for bounded uncertainty for two reasons. First, users have been trained to interpret them as probabilistic bounds on an unbounded distribution of possible values. Second, since error bars are typically used in conjunction with an estimated exact value, the existence of a single most likely value is strongly implied.

Visualizations should clearly differentiate between the two forms of uncertainty, making it obvious whether the uncertainty is statistical or bounded in addition to conveying the degree of uncertainty. Therefore, we advocate the use of two distinct techniques for the two forms of uncertainty. To convey statistical uncertainty, it is appropriate to display the most likely value along with error bars or other glyphs as in [PWL97]. To convey bounded uncertainty, we advocate a systematic technique based on widening the boundaries and positions of graphical elements and rendering the uncertain region in fuzzy ink. We show how to apply this technique, which we call *ambiguation*, to common displays of abstract charts and graphs. Interestingly, it is not always possible to show the exact degree of uncertainty, and in some cases it can only be displayed approximately.

The remainder of this paper is structured as follows. We begin by discussing related work in Section 2. We then formally define the two forms of uncertainty in Section 3. Then, in Section 4 we describe our systematic approach to conveying the presence, form, and degree of uncertainty.

*Research supported by a National Science Foundation graduate research fellowship.

2 Related Work

In certain visualization scenarios, data may be unavailable for display or even purposefully omitted for a variety of possible reasons, giving rise to uncertainty. The importance of visually informing the user of the absence of data has been identified [WO98] and techniques for doing so have been proposed in, *e.g.*, Clouds [HAC+99] and Restorer [TCS94]. We focus on a different type of uncertainty where all the data is present but precise values are not known.

Numerous ways to convey the degree of uncertainty in data using overlayed annotations and glyphs have been proposed, as in, *e.g.*, [PWL97]. Another approach is to make the positions of grid lines used for positional reference ambiguous [CR00]. Uncertainty can also be indicated by adjusting the color, hue, transparency, etc. of graphical features as in, *e.g.*, [DK97, Mac92, vdWvdGG98]. Some techniques for conveying uncertainty by widening the boundaries of graphical elements have also been proposed. For example, in [WSF+96], the degree of uncertainty in the angle of rotation of vectors is encoded in the width of the vector arrows. Also, [PWL97] proposes varying the thickness of three-dimensional surfaces to indicate the degree of uncertainty.

To our knowledge, however, none have focused on accurately and unambiguously conveying not only the presence and degree but also the form of uncertainty in data, as we do. We also believe that our work is the first to establish systematic methods for conveying bounded uncertainty by widening the boundaries and positions of graphical elements in abstract charts and graphs. The approach in [FWR99] for displaying cluster densities gives a visual appearance similar to our ambiguated line charts (discussed later) but serves a different purpose.

3 Forms of Uncertainty

In this paper we consider two commonplace forms of uncertainty, as described in [TK94], [PWL97], and elsewhere. Consider a numeric data object O whose exact value V is not known with certainty. There are two predominant ways in which partial knowledge about the possible values of V can be represented: *statistical uncertainty* and *bounded uncertainty*. Under statistical uncertainty, the uncertain value of a data object can be represented in a number of ways, depending on the statistical model. In one common case, when errors follow a normal distribution, the uncertain value of a data object can be represented by a three-tuple $\langle E, D, P\rangle$ of real numbers, where $D \geq 0$ and $P \in (0, 1]$. Here, E is an estimate that represents the most likely candidate for the unknown value V, and P is the probability that V lies in the confidence interval $[E - D, E + D]$. Typically, P is fixed at, say, $P = 0.95$, and D is chosen so that the value V lies inside the confidence interval $[E - D, E + D]$ with probability P. Under bounded uncertainty, there is some numeric interval $[L, H]$ that is guaranteed to contain the exact value V, *i.e.*, $L \leq V \leq H$. Under bounded uncertainty, the probability that V is outside the interval is zero, but, unlike with statistical uncertainty, no assumptions can be made about the probability distribution of possible values inside the interval.

Both forms of uncertainty commonly occur in scientific and other applications [TK94]. For example, bounded uncertainty can occur when measurements are taken using a device having an unknown degree of imprecision that lies within known bounds. Statistical uncertainty can occur, for example, when single or repeated measurements are taken in conditions exhibiting experimental variability, often resulting in an unbounded probability distribution over possible values featuring a central peak. Both bounded and statistical uncertainty can also occur in emerging data delivery paradigms that intentionally introduce uncertainty for performance reasons, *e.g.*, [HAC+99, OW02]. In these paradigms there is often the opportunity to adjust the uncertainty levels interactively, unlike with traditional sources of uncertainty. In the extended version of this paper [OM02], we discuss interactive data delivery techniques that exhibit these properties.

4 Representing Uncertain Data Visually

Having described the two common forms of uncertainty and some ways they can occur, we are now ready to discuss ways to represent uncertain data visually. In most abstract charts and graphs, data values are graphically encoded either in the positions of graphical elements, as in a scatterplot, or in the extent (size) of elements along one or more dimensions, as in a bar chart. When the underlying data is uncertain, we believe it is appropriate to clearly indicate not only the presence and degree but also the form of uncertainty. As described in Section 3, statistical and bounded uncertainty encode two dramatically different distributions of potential values. Due to this key difference, using the same display technique to represent both forms of uncertainty could mislead the user. Instead, we advocate two alternative methods for conveying uncertainty in the positions or extents of graphical representations of data: *error bars* for statistical uncertainty and *ambiguation* for bounded uncertainty. We begin by describing these general techniques and then show how they can be applied to some common types of charts and graphs.

4.1 Error Bars

Error bars and their variants have been well studied as a suitable means to convey statistical uncertainty [Cle85, Tuf01, Tuk77]. For each uncertain data value to be represented visually, the idea is to use the normal display technique to render the estimate E in place of the unknown exact value V. Error bars are then added to indicate uncertainty in the position or boundary location in proportion to the size of the confidence interval $[E - D, E + D]$. Some standard uses of error bars are illustrated in the upper left quadrant of Figure 1. When uncertainty occurs in bounded rather than statistical form, it is important to avoid the use of error bars since the accepted interpretation implies a potentially unbounded distribution extending beyond the error bars. Even worse, rendering an exact estimate using the normal display technique strongly implies the existence of a most likely value E, but in bounded uncertainty no most likely value can be assumed.

4.2 Ambiguation

To convey the presence and degree of bounded uncertainty, we propose the use of a technique we call ambiguation. The main idea behind ambiguation when uncertain data is encoded in the extent of a graphical element is to widen the boundary to suggest a range of possible boundary locations and therefore a range of possible extents. The ambiguous region between possible boundaries can be drawn as graphical fuzz, giving an effect that resembles

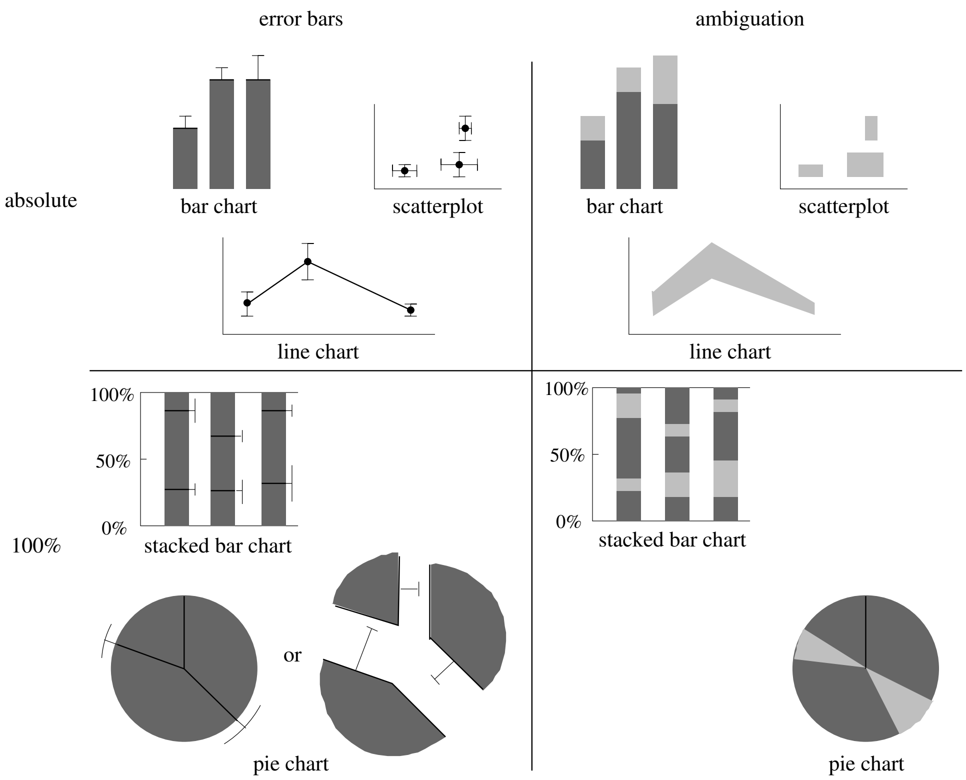

Figure 1: Error bars and ambiguation applied to some common chart types.

ink smearing. A straightforward application of this technique is illustrated in the ambiguated bar chart in the upper right quadrant of Figure 1. To indicate positional uncertainty, rather than drawing a crisp representation of the graphical element at a particular position, the representation is elongated in one or more directions and drawn using fuzz. A simple application of this technique is illustrated in the ambiguated scatterplot in the upper right quadrant of Figure 1.

Other variations of boundary or position ambiguation may be possible, but the necessary feature is that no particular estimate or most likely value should be indicated. Rather, the entire range of possible values for the boundary or position of the graphical element should be presented with equal weight. This key characteristic is in contrast with error bars and other approaches such as fuzzygrams and gradient range symbols [Har99] that emphasize a known probability distribution over data values.

4.3 Discussion

The complementary use of error bars and ambiguation makes the presence, degree, and form of uncertainty clear. First, these techniques make it easy to identify the specific data values that are uncertain by suggesting imprecision in the graphical property (position or boundary location) in which the values are encoded. For bounded uncertainty, the position or boundary is made ambiguous using fuzzy ink, and for statistical uncertainty, error bars are added to visually suggest the possibility of a shift in position or boundary location. Second, these techniques allow the degree of uncertainty to be read in a straightforward manner using the same scale used to interpret the data itself. Finally, the use of two visually distinct techniques makes it clear which of the two forms of uncertainty is present, and each technique conveys the properties of the form of uncertainty it represents.

Ambiguation and error bars work well when data is encoded as the position or extent of graphical elements. Coping with displays that use other graphical attributes such as color and texture to encode data is left as a topic for future work. In the absence of analogous techniques for other graphical attributes, when uncertainty is present it is desirable to only use charts and graphs that encode data using position and extent alone so the presence, degree, and form of uncertainty can be clearly and unambiguously depicted.

4.4 Application to Common Chart Types

Figure 1 illustrates how error bars and ambiguation can be applied to some common chart types (exhaustive illustration on all known chart types is omitted for brevity). While these techniques are general and can be applied to a broad range of displays that use position and extent to encode data, we focus on abstract charts and graphs, which can be classified into two categories: *absolute displays* and *100% displays*. In absolute displays, each data value is given a graphical representation whose extent or position is plotted on an absolute scale. Examples of absolute displays include simple bar charts (which encode data in the upper boundaries of bars), scatterplots (which encode data in the positions of points), and line graphs (which encode data in the positions of points and lines). It is generally straightforward to add error bars or apply ambiguation to boundaries and positions in absolute displays such as those displayed in the top half of Figure 1.[1]

In 100% displays, the scale ranges from 0% to 100%, and n values $V_1, V_2, \ldots, V_n$ are plotted on this relative scale. Each value V_i is plotted as a graphical element whose size is proportional to the fraction $\frac{V_i}{\sum_{1 \leq j \leq n} V_j}$ of the total over all n values. Examples of 100% displays include stacked bar charts and pie charts. Indicating uncertainty in 100% displays is more challenging than doing so in absolute displays. In 100% displays, the graphical elements usually contact each other directly, so the boundary between two elements indicates the difference between them in terms of relative contribution to the total. To inform the user of statistical uncertainty in the locations of these boundaries, error bars can be drawn adjacent to the boundaries. Alternatively, for pie charts the wedges can be separated, leaving space for error bars extending directly from the boundaries between wedges. The lower left quadrant of Figure 1 illustrates these techniques. Bounded uncertainty can be indicated by inserting an ambiguous region of fuzzy ink between each pair of elements whose shared boundary is uncertain, as illustrated in the lower right quadrant of Figure 1. It turns out that determining the sizes to use for the fuzzy and solid regions in an ambiguated 100% display is not trivial because each region of fuzz shares a border with two solid data regions. In fact, it is not always possible to show the exact degree of uncertainty, and in some cases it can only be displayed approximately.

In the extended version of this paper [OM02], we specify an algorithm that approximates the degree of uncertainty to make it displayable while minimizing the overall loss in accuracy. In addition, in [OM02] we consider new data delivery paradigms that offer mechanisms for interactive control over uncertainty levels, but whose use may result in hidden side effects. We propose interfaces that offer control of uncertainty levels to the user in ways that encourage careful use of these facilities.

References

[Cle85] W. S. Cleveland. *The Elements of Graphing Data*. Wadsworth, 1985.

[CR00] A. Cedilnik and P. Rheingans. Procedural annotation of uncertain information. In *Proceedings of the IEEE Visualization Conference*, pages 77–84, Salt Lake City, Utah, October 2000.

[DK97] T. J. Davis and C. P. Keller. Modelling and visualizing multiple spatial uncertainties. *Computers and Geosciences*, 23(4):397–408, 1997.

[FWR99] Y. Fua, M. O. Ward, and E. A. Rundensteiner. Navigating hierarchies with structure-based brushes. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 58–64, San Francisco, California, October 1999.

[HAC+99] J. M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, and P. Haas. Interactive data analysis with CONTROL. *IEEE Computer*, August 1999.

[Har99] R. L. Harris. *Information Graphics*. Oxford University Press, 1999.

[Mac92] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspective*, (13):10–19, 1992.

[OM02] C. Olston and J. Mackinlay. Visualizing data with bounded uncertainty. Technical report, Stanford University Computer Science Department, 2002. http://dbpubs.stanford.edu/pub/2002-20.

[OW02] C. Olston and J. Widom. Approximate caching for continuous queries over distributed data sources. Technical report, Stanford University Computer Science Department, 2002. http://dbpubs.stanford.edu/pub/2002-8.

[PWL97] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, pages 370–390, November 1997.

[TCS94] R. Twiddy, J. Cavallo, and S. M. Shiri. Restorer: A visualization technique for handling missing data. In *Proceedings of the IEEE Visualization Conference*, pages 212–216, Washington, D.C., October 1994.

[TK94] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of nist measurement results. Technical report, National Institute of Standards and Technology Note 1297, Gaithersburg, Maryland, 1994.

[Tuf01] E. R. Tufte. *The Visual Display of Quantitative Information, Second Edition*. Graphics Press, 2001.

[Tuk77] J. W. Tukey. *Exploratory Data Analysis*. Addison Wesley, 1977.

[vdWvdGG98] F. J. M. van der Wel, L. C. van der Gaag, and B. G. H. Gorte. Visual exploration of uncertainty in remote-sensing classification. *Computers and Geosciences*, 24(4):335–343, 1998.

[WO98] A. Woodruff and C. Olston. Iconification and omission in information exploration. In *SIGCHI '98 Workshop on Innovation and Evaluation in Information Exploration Interfaces*, Los Angeles, CA, April 1998. http://www.fxpal.com/ConferencesWorkshops/CHI98IE/submissions/woodruff.htm.

[WSF+96] C. M. Wittenbrink, E. Saxon, J. J. Furman, A. T. Pang, and S. K. Lodha. Glyphs for visualizing uncertainty in environmental vector fields. *IEEE Transactions on Visualization and Computer Graphics*, pages 266–279, September 1996.

[1] Displays such as stacked bar charts that are not normalized to sum to 100% are problematic when used in conjunction with these uncertainty indicators because the interpretation can be ambiguous. For example, in an absolute stacked bar chart, an error bar or a fuzzy region appearing at the top of the stack can be interpreted either as uncertainty in the topmost element or as uncertainty in the overall height of the stack (the sum over all elements).