Eurographics/ IEEE-VGTC Symposium on Visualization 2010
G. Melançon, T. Munzner, and D. Weiskopf
(Guest Editors)

Volume 29 (2010), Number 3

Visualizing Summary Statistics and Uncertainty

K. Potter[1], J. Kniss[2], R. Riesenfeld[3], and C.R. Johnson[1]

[1]Scientific Computing and Imaging Institute, University of Utah
[2]Department of Computer Science, University of New Mexico
[3]School of Computing, University of Utah

Abstract

The graphical depiction of uncertainty information is emerging as a problem of great importance. Scientific data sets are not considered complete without indications of error, accuracy, or levels of confidence. The visual portrayal of this information is a challenging task. This work takes inspiration from graphical data analysis to create visual representations that show not only the data value, but also important characteristics of the data including uncertainty. The canonical box plot is reexamined and a new hybrid summary plot is presented that incorporates a collection of descriptive statistics to highlight salient features of the data. Additionally, we present an extension of the summary plot to two dimensional distributions. Finally, a use-case of these new plots is presented, demonstrating their ability to present high-level overviews as well as detailed insight into the salient features of the underlying data distribution.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques

1. Introduction

As computational power, memory limits, and bandwidth have inexorably increased, so has the corresponding size and complexity of the data sets generated by scientists. Because of the reduction of hardware limitations, simulations can be run at higher resolutions, for longer amounts of time, using more sophisticated numerical models. We can compute more exhaustively, store more abundantly, and access data more rapidly, all of which leads researchers to create more complex systems to increase the accuracy and reduce the error in scientific simulations.

As data becomes increasingly large and complex, visualization and data analysis techniques are required that not only address issues of large scale data, but also allow scientists to better understand the processes that produce the data and the nuances of the resulting data sets. Uncertainty, in the form of confidence, variability, and error, as well as model bias and trends, is regularly included within data sets and is used to express descriptive, qualitative characteristics of the data. Because uncertainty is crucial in understanding the reliability of information and thus in objectives such as decision making, its absence can lead to misrepresentations and incorrect conclusions. Too often, traditional visualization approaches overlook available uncertainty information [JS03, Joh04]. As the importance of visualizing these large, complex data sets grows, the actual task of visualizing them becomes more complicated; incorporating the additional data parameter of uncertainty into the visualizations becomes even less straightforward. Difficulties in applying preexisting methods, additional visual clutter, and the lack of obvious visualization techniques leave uncertainty visualization an unsolved problem.

The goal of this work is to create a summary plot that incorporates higher order descriptive statistics to concisely present data with uncertainty information. This work takes inspiration from the visual devices used in exploratory data analysis and extends their application to uncertainty visualization. The statistical measures often used to describe uncertainty are similar to measures conveyed in graphical devices such as the histogram and box plot. This research investigates the creation of the *summary plot*, which combines the box plot, histogram, a plot of the central moments (mean, standard deviation, etc.), and distribution fitting. The box plot has a canonical feel; the "signature" of the plot is easily recognizable and does not need much explanation to allow for a full understanding. The focus of this work is to create a

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

summary plot that similarly incorporates higher-order information, allowing for the quick identification of characteristic features. This higher-order signature provides at-a-glance recognition of variations from normal and allows easy comparison of data distributions in detail. In addition, a 2D extension of the summary plot is presented, which provides for the comparison of correlated data. Finally, an exemplary application of the method demonstrates the ability of the summary plot to highlight variabilities in a data set.

2. Background

Understanding data sets is an essential part of the scientific process. However, discerning the significance of data by looking only at numerical values is a formidable task. Descriptive statistics are a quick and concise way to extract the important characteristics of a data set by summarizing the distribution through a small set of parameters. Measures of central tendency, variation, and quantiles are typically used for this purpose. The main goal of descriptive statistics is to quickly describe the characteristics of the underlying distribution of a data set through a simplified set of values. Often these parameters provide insights into the data that would otherwise be hidden. In addition, data summaries facilitate the presentation of large scale data and comparison of multiple data sets.

Creating graphics for data presentation is a difficult task involving decisions not only about data display, but also about data interpretation. The graphic is often intended to show specific characteristics of the data, and the presentation style should make these characteristics clear. Numerous sources outline design practices for effective data visualization [CCKT83,Cle94,Tuf83,Wil99b]. These references not only direct the researcher towards the “correct” graphical technique for specific data types, but also describe how a visualization may be interpreted by the viewer and suggest methodologies to influence this interpretation.

One of the most common approaches to graphing summary statistics is the box plot [Hae48,FHI89,Spe52,Tuk77], which is the standard technique for presenting the *5-number summary*, consisting of the minimum and maximum range values, the upper and lower quartiles, and the median, as demonstrated in Figure 1(a). This collection of values quickly summarizes the distribution of a data set, including range and expected value, and provides a straightforward way to compare data sets. Figure 1(b-d) shows various visual modifications on the box plot, and surveys on its introduction and evolution can be found in [CM05,Pot06].

The box plot is often adapted to include information about the underlying distribution, as demonstrated in Figure 1(e-j). The most common modifications add density information, typically through changes to the sides of the plot. The hist plot [Ben88] extends the width of the cross bars at the quartiles and median to express density at these three locations. The vase plot [Ben88] instead varies the “box” continuously to reflect the density at each point in the interquartile range. Similarly, the box-percentile plot [EB03] and violin plot [HN98] show density information for the entire range of the data set. Density can also be shown by adding dot plots [Wil99a], which graph data samples using a circular symbol. The sectioned density plot [CC06] completely reconstructs the box plot by creating rectangles whose colors and size indicate cumulative density, and placement express the location of the quartiles. Sample size and confidence levels can be expressed through changing or notching the width of the plot [MTL78] or by using dot-box plots, which overlay dot plots [Wil99b]. Other descriptors, such as skew and modality, can be added by modifying the width of the median line [MTL78], thickening the quartile lines [CM05], or adding beam and fulcrum displays [DT00]. Multivariate extensions of the box plot expand it into two dimensions [BG87,GI92,RRT99,Ton05].

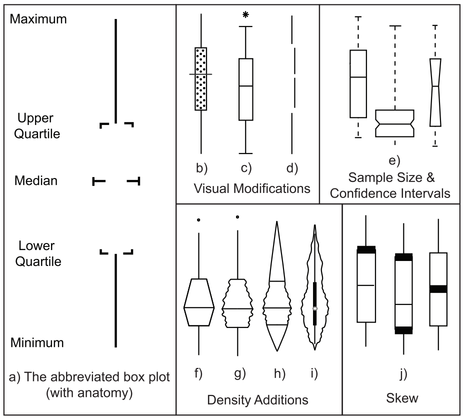


Figure 1: *Variations on the box plot. a) Abbreviated box plot. b) Range plot [Spe52]. c) Box plot [Tuk77]. d) Interquartile plot [Tuf83]. e) Variable width and notched box plots [MTL78] expressing sample sizes and confidence levels. f) Hist plot [Ben88] g) Vase plot [Ben88] h) Box-percentile plot [EB03]. i) Violin plot [HN98]. j) Skew and modality plots [CM05].*

3. The Summary Plot

The hybrid box plot we are introducing can be more formally titled the *summary plot*. This display includes not only the quartile information present in the form of a modified box plot, but also a collection of descriptive statistics and density information. As shown in Figure 2, we use an abbreviated form of the traditional box plot to convey the 5-number summary and a symmetrically drawn histogram to show density information. While this technique is similar to that of the violin plot [HN98], we have extended it to include minimum and maximum rather than truncating extreme values, and incorporated a colormapped histogram to further aid in understanding. Descriptive statistics, in the form of mean, stan-

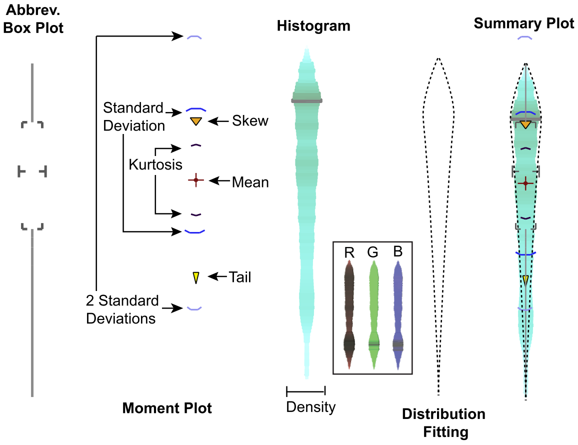


Figure 2: *Anatomy of the summary plot. The abbreviated box plot displays the range of the data distribution. The moment plot shows higher order statistics which describe feature characteristics. The histogram estimates the density of the distribution and is displayed using a symmetric display and a redundant colormap. Distribution fitting allows the user to compare the data against well-known distributions.*

dard deviation, and higher-order moments, are expressed as glyphs with the design of each reflecting the semantic meaning of the statistic. Finally, distribution fitting capabilities are added to allow the user to compare against and find a best fit from a library of well-known distributions.

3.1. The Abbreviated Box Plot

As discussed, the traditional approach to presenting summary information is through the box plot, which has been refined numerous times in efforts to maximize the ratio of information to ink consumption and improve aesthetics. We have chosen to refine the plot further, as shown on the left of Figure 2. Our plot builds on Tukey's box plot [Tuk77] (Figure 1(c)) with a few important distinctions. The first modification removes the sides of the interquartile box. This not only reduces the visual real estate of the plot, but also removes possible assumptions incurred from the sides of the box about the density of the distribution. The prevalence of using the sides of the plot to indicate density is due to the visual metaphor created by the box itself. Since 50% of the data samples lie within the box, it is easy to assume a density distribution that resembles the plot itself, with the highest densities falling close to the median. However, this restricts the plot to normal or Gaussian-like distributions, which is not always the case. Often, the mode (or most frequently occuring sample value) lies outside of the interquartile range, which is only evident when the box plot is combined with a density display. Furthermore, outliers are not removed from the plot. While this choice may stretch the plot to extreme values, we prefer to express the entire range of the data set within the plot, rather than add additional glyphs to indicate outliers which will increase the visual complexity of the plot.

3.2. The Histogram

Density information is added to the summary plot as a histogram, which is displayed using quadrilaterals whose widths are varied with the density at each bin location. The colormap used for the histogram was designed to be both redundant and non-intrusive. Each color channel uses a distinct mapping, which when combined, clearly emphasizes areas of high density without overpowering the plot with color. The color channel is defined as follows: red is normalized log density, green is normalized density, and blue is normalized linear density. Each color channel can be seen in Figure 2 (inset). The distinction between the maps for the individual channels is subtle and intended to encode the density information in a manner that is reiterative, aesthetically pleasing, and subdued so as to act as a backdrop for the more saturated color scheme used for other plots.

3.3. Moments

The following is a list of the equations used to calculate the various moments, and the notation that will be used throughout the paper:

Given a data set $\{x_i\}_{i=1}^{N}$, we define the following quantities:

kth Central Moments:	$\mu_k \simeq \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_1)^k$
Mean:	$\mu_1 \simeq \frac{1}{N}\sum_{i=1}^{N}x_i$
Variance:	$\mu_2 \simeq \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_1)^2$
Standard Deviation:	$\sigma = \sqrt{\mu_2}$
Skew:	$\gamma = \frac{\mu_3}{\sigma^3}$
Kurtosis:	$\kappa = \frac{\mu_4}{\sigma^4}$
Excess Kurtosis:	$\kappa_e = \kappa - 3$
Tailing:	$\tau = \frac{\mu_5}{\sigma^5}$

where N is the number of data samples.

The moments of a distribution are statistical measures of feature characteristics. The main distinction between the summaries presented by the box and the moment plots is that the quartiles give information about the location and variation changes in the data, while moments express descriptive characteristics of the look of the distribution such as "peakedness." These measures not only highlight uncertainty through the standard deviation, but also give indications as to where the variation in the data set stems, such as subsets of the data diverging from the mean.

One of the drawbacks of using only a box plot to summarize a distribution is that multiple, distinct distributions can have the same box plot signature. For example, one may come across two distributions, one unimodal (having one data value occurring most frequently) and the other multimodal (multiple most frequent values), having identical quartiles and thus indistinguishable box plot signatures. Adding moment information exposes differences between distributions and allows for the expression of non-Gaussian distributions, while maintaining the simplistic nature of the original box plot.

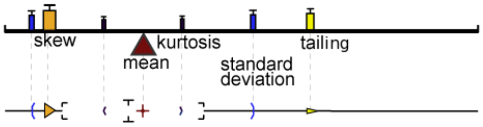


Figure 3: *Moment arm abstraction from which we designed the moment plot. Using the balance bean metaphor, each glyph is placed so as to stabilize the weight on the beam.*

The use of moments in physics provides valuable insight into how moments express characteristics of a data distribution (Figure 3). In this example, a beam is placed on a fulcrum, the position of which is dictated by the mean [BE92]. The moments can then be thought of as weights used to balance the beam, each moment having a specific role in dynamically balancing the system. While this approach is not meant to be a physically-based explanation of moments, those unfamiliar with the role of moments in statistics may find this abstraction helpful.

3.3.1. Mean and Standard Deviation

The most familiar and frequently used moments are mean and variance (the first and second moment). The average of the data samples is an unbiased estimator of the mean of the underlying distribution, or the expected value of a random variable. Variance is a measure of the dispersion of the data, indicating the distance a random variable is likely to fall from the expected value. Standard deviation is simply the square root of variance. For the summary plots, we use only mean and standard deviation, as standard deviation is derived from variance and is typically used as a measure of uncertainty.

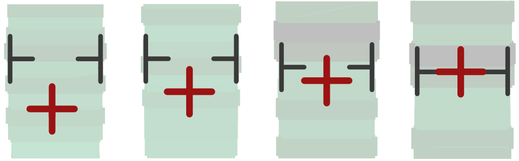


Figure 4: *The mean is represented by a red cross and the median by dark grey lines on the left and right. The mean and median glyphs align when the values are equal, thus easing visual comparison to normal distributions.*

The addition of mean and standard deviation to the summary plot is straightforward. The mean is rendered as a dark red cross. The width of the lines making up the cross are constructed so that when the mean and median are displayed at the same location, the glyphs coincide and form a straight line across the plot. This emphasizes symmetrical distributions and quickly reveals when a distribution varies from normal. A close up of this can be seen in Figure 4.

Standard deviation, like all even moments, is rendered as two glyphs on the plot. Two blue curved lines are placed on either side of the mean to express the average variation from the mean. The glyphs are placed at mean ± standard deviation and mean ± 2×standard deviation. This placement allows the user to easily see where the majority of the data lies, as well as to identify samples outside two standard deviations, typically referred to as extrema.

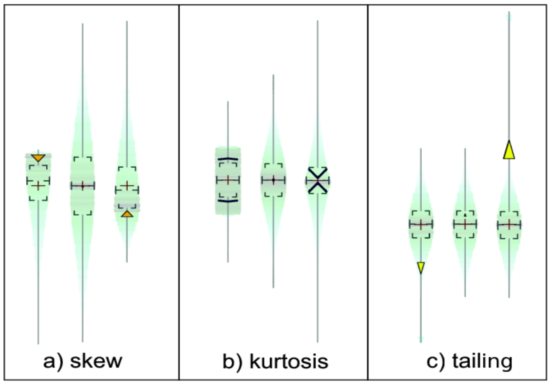


Figure 5: *Glyphs for the higher-order central moments. Each triplet of distributions shows negative, close to zero, and positive values for the respective moment. Each higher-order moment is relative to the moments of a Gaussian distribution, which is the central distribution in each set.*

3.3.2. Skew

Skew is a measure of the asymmetry of a distribution, or the extent to which the data is pushed to one side or the other. Figure 5(a) shows three distributions in which skew varies from negative to positive. Based on the balance beam abstraction (see Figure 3), we use a triangle to denote skew in the summary plot. The triangle is scaled by the absolute value of skew, clamped so that very large skew values do not extend the glyph beyond the the boundaries of the summary plot. The triangle placement rests the glyph on the side of the distribution with the highest density, pointing at the end with the longest tail. Mathematically, we calculate the placement of the skew glyph by first finding skew (γ) as defined above, and placing the glyph $-\gamma$ distance away from the mean, with the apex of the triangle pointing toward the tail of the distribution. Thus, the placement of the skew glyph indicates on which side of the mean the largest spatial grouping of samples lies.

3.3.3. Kurtosis

Kurtosis is a measure of how peaked or flat topped a distribution is compared to a normal distribution. Excess kurtosis is the standard kurtosis measure normalized by the kurtosis of a Gaussian. Figure 5(b) shows three distributions with varying kurtosis, where a flat, box-like distribution can be seen on the left. This type of distribution has large, negative kurtosis (*i.e.*, $\kappa_e < 0$) and is called *platykurtic*. Moving right, the kurtosis value gets very close to a *mesokurtic* (normal) distribution (*i.e.*, $\kappa_e = 0$) and then to a highly peaked, *leptokurtic* (*i.e.*, $\kappa_e > 0$) distribution.

The glyphs chosen to represent kurtosis reflect the aforementioned categories of kurtosis. The glyphs are rendered using a deep purple color and are scaled so that their size reflects their magnitude away from 0 (excess kurtosis). To distinguish between flat and peaked, the glyphs assume a flat or sharp shape depending on the sign of kurtosis. For a highly positive value, the glyph is very pointy; the more negative the kurtosis value, the flatter the glyph.

3.3.4. Tailing

The final moment added to the summary plot is what we refer to as tailing, which is based on the fifth central moment, μ_5. The quantity is sensitive to distribution asymmetry farther away from the mean when compared to the skew. Tailing will have a high magnitude when there are additional modes in the distribution or strong outliers. Like skew, tailing is rendered as a triangle pointing in the direction of asymmetry. However, unlike skew, the glyph is rendered on the same side of the mean as its sign. The size and sharpness of the arrowhead is dependent on the tailing quantity, and the visual effect of this glyph indicates when there is a significant number of samples very far from the mean. Figure 5(c) shows a set of distributions with tailing values varying from very negative to very positive. Upon close inspection, one can see a cluster of outliers in the rightmost distribution, which is indicated by the large size of the glyph.

3.4. Distribution Fitting

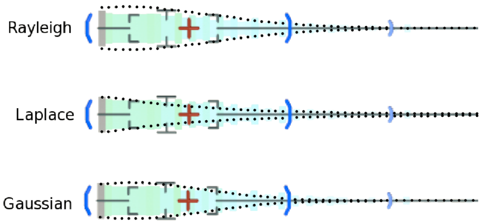


Figure 6: *The results of fitting 3 canonical distributions to a single data set are shown as dotted lines on either side of the plot.*

Understanding the characteristics of a particular data set is often less interesting than determining the canonical distribution that best fits the data because the feature characteristics of the canonical distributions, such as the Gaussian, are well known. The final element of the summary plot is a distribution fit plot, which represents either a best-fitting distribution or a user-chosen distribution. The user is provided with a library of common distributions including Gaussian, Uniform, Poisson, Rayleigh, Laplace, and others, as well as the fitting of multiple Gaussians and asymmetric distributions. The fit distribution is displayed symmetrically as a dotted line showing the density along the axis, as seen in Figure 6. In addition, any distribution can be used as a learning set, allowing the user to quickly identify data sets that resemble specific distributions, as well as to explore relationships between distributions.

4. Joint 2D Summaries

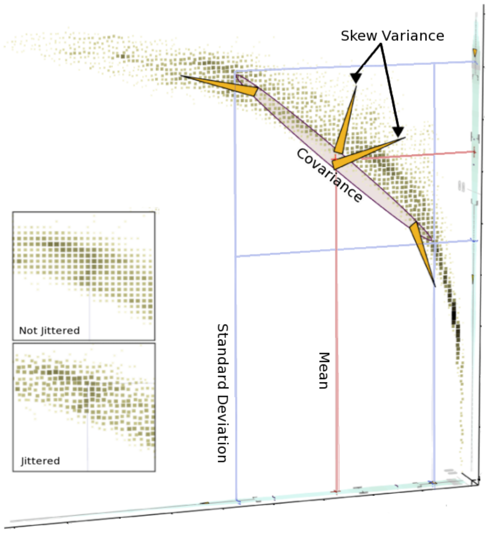


Figure 7: *Joint summary for two 1D categorical data sets. The red and blue lines show the joint mean and standard deviation, respectively. The joint histogram is shown as colormapped, jittered quadrilaterals emphasizing where both data sets express density. Covariance and skew variance are shown as glyphs centrally located within the area created by the lines of standard deviation. These glyphs show how much the two data sets vary together, as well as how much they are skewed in the same direction.*

In addition to a statistical summary for a 1D categorical data set, users require methods for comparing multiple, correlated data sets to understand how data values are related. In this section, we explore methods for summarizing categorical data with pairs of values associated with each sample. Figure 7 shows the joint summary plot for two 1D data sets. The joint summary places 1D summary plots for each data set perpendicularly to orient the viewer. Joint mean and standard deviations, a joint histogram, and a reduced higher-order moment plot are added, providing a display which shows the relationship between correlated data sets.

Note that we drop the summary of cumulants for higher-dimensional distributions. We do feel, however, that a box plot type summary is important even in higher dimensions. Thus we refer to a generalization of the box plot, known as the bag plot [RRT99]. Unfortunately, the bag plot approach does not necessarily have the same correspondence to cumulant distributions as does the box plot. It is a suitable approximation for many applications, but we defer discussion of multivariate cumulant summaries to future work.

4.1. Joint Mean, Standard Deviation, and Density

The first measures of correlation that we add to the display are joint mean, standard deviation and density, which are shown in Figure 7. The display of joint mean and standard deviation uses lines that connect the mean and standard deviation of one distribution to the corresponding values in the other, taking the colors of these measures from the summary plot. A joint histogram is used to display the density of a set of samples drawn from a 2D distribution. Our system displays the joint histogram by rendering a quadrilateral that is both colormapped and scaled to show the density at each bin of the 2D distribution. This is meant to be reminiscent of a scatter plot of the joint density for the correlated distributions. The inset of Figure 7 shows how jittering can alleviate aliasing artifacts that occur when multiple joint summaries are presented together, as shown in Figure 10.

4.2. Covariance and Skew Variance

For multivariate distributions, the covariance matrix is the analogue of variance in 1D distributions. The covariance of two data sets, $\{x_i\}_{i=1}^N, \{x_j\}_{j=1}^N$ can be defined,

$$V_{ij} = \frac{1}{N}\sum_{k=1}^{N}(x_{i_k} - \mu_i)(x_{j_k} - \mu_j)$$

where μ_i and μ_j are the means for each data set. Covariance is a measure of how the two data sets vary in relation to each other. For our presentations, the covariance matrix is used to transform a unit disk so that the visual stretch of the disk relates to the covariance of the data sets. Since we are interested in a multivariate analogue of standard deviation, we scale the covariance ellipse-disk glyph as follows: $\text{scale} = \frac{\sqrt{\text{ev}_{\text{max}}}}{\text{ev}_{\text{max}}}$, where ev_{max} is the maximum eigenvalue of the covariance matrix.

Just as covariance is the analogue of variance, higher-order multivariate moments can also be described with matrices. The so called "skew variance" of two data sets, $\{x_i\}_{i=1}^N, \{x_j\}_{j=1}^N$ can be expressed by two matrices, $V_{i^2 j^1}$ and $V_{i^1 j^2}$ where

$$V_{i^m j^n} = \frac{1}{N}\sum_{k=1}^{N}(x_{i_k} - \mu_i)^m (x_{j_k} - \mu_j)^n$$

In general, these matrices are neither symmetric nor positive definite. Skew variance is visualized using four sharp arrows pointing in the direction of the skew located at the endpoints of the covariance eigenvectors. The directions of the skew variance arrows are defined by the column vectors of $V_{i^2 j^1}$ and $V_{i^1 j^2}$. As with covariance, skew-variance visualizations are scaled: $\text{scale} = \frac{\sqrt[3]{\text{ev}_{\text{max}}}}{\text{ev}_{\text{max}}}$, where ev_{max} is the maximum eigenvalue of the skew-variance matrix.

The use of skew-variance glyphs in 2D (or higher dimensional) distributions is important, since joint distributions can be very asymmetrical even when their 1D distributions are symmetrical. While the covariance ellipse indicates the overall trend of the joint distribution, it gives no indication that the majority of the distribution's density is outside the ellipse. The skew variance glyphs indicate the strong asymmetry of this distribution. When multiple 2D distributions are combined, as seen in Figure 10, top inset, the moment glyphs allow the user to visually identify each category. Without them, the individual joint histograms would be difficult to separate.

5. Short-Range Ensemble Forecasts

We demonstrate the use of summary plots on NOAA's Short-Range Ensemble Forecast (SREF), a data set publicly available from the National Centers for Environmental Prediction's (NCEP) Environmental Modeling Center and Short-Range Ensemble Forecasting Project [NCE]. The main challenges in using this data stem from its size and complexity. The SREF ensemble contains 21 *members* comprising four numerical models, each run with various parameter perturbations. A single member contains 624 state variables predicted at each of the 24,00 points of the regular grid across North America and is run out to 87 forecast hours. Each full run of the SREF ensemble contains 36GB of data, resulting in 108GB every day. Colormaps of the mean and standard deviation can be seen in Figure 8, left and center.

To apply the summary plots on the SREF data, one must decide which part of the data is interesting; generating a summary plot for every data point is not feasible for display. This can be done by allowing the user to select regions of interest, or automatically selecting as shown in Figure 9, top, right. Here, we use the k-medoids clustering algorithm [Bis06] to select regions of the domain that exemplify the variation across the region. In this case, we are looking for areas of high variation in order to understand locations of high uncertainty which indicates where the mean of the data is an unreliable estimation of the outcome.

Figure 9 shows the results of using the summary plots on the representative cluster points of the SREF data. To ease visual interpretation within this publication, the summary plots for each cluster location are extracted from the plot display and enlarged. The positions of each summary plot within the plot axis are indicated by a box plot which also clearly shows the range of the data. In practise, the user can zoom in and out of the plot to more closely investigate the summary plots.

On first glance the cluster positions with high uncertainty standout, particularly points 10, 11, and 12. This is clear from the length of the plots, as well as the strength of the standard deviation glyphs. More interestingly, plot 11 has strong groupings within the density display, indicating a disagreement in the predicted outcome. That is, some subset of the ensemble mainly predict one particular outcome, while another group predicts another, different outcome. This is in

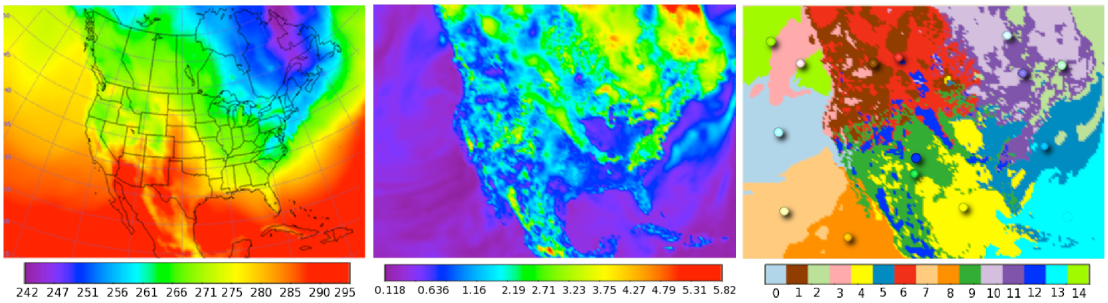


Figure 8: *Temperature at 2M above ground at valid forecast hour 27. Color refers to the mean (leftmost) and standard deviation (center) of the ensemble computed at each grid point. Rightmost, results of k-medoids clustering algorithm [Bis06] on the temperature data. The domain is colormapped based cluster membership.*

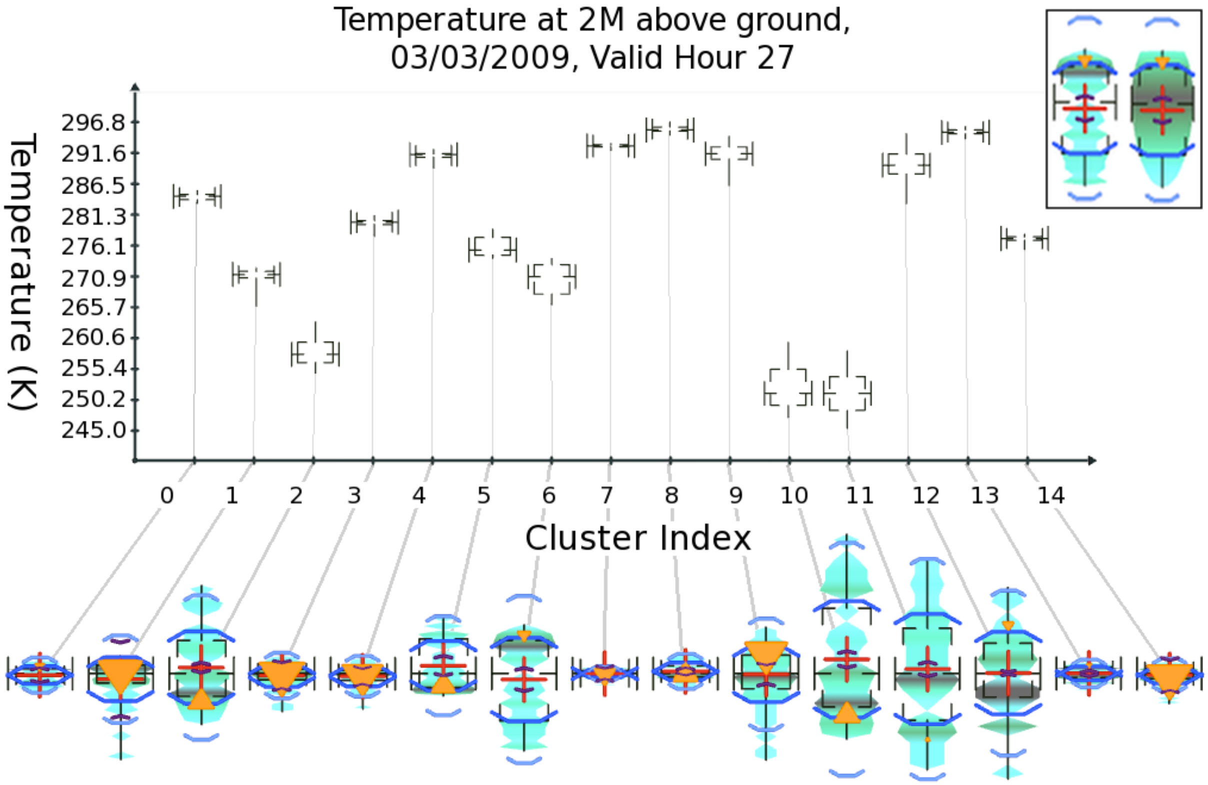
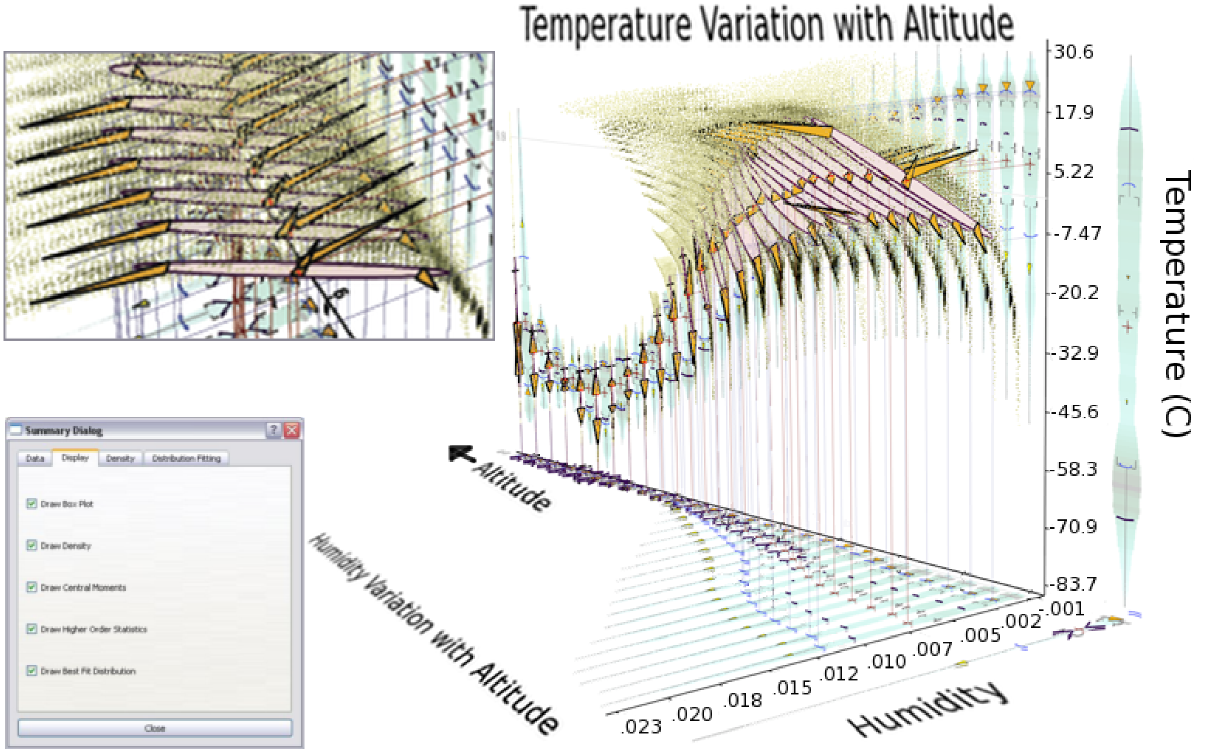


Figure 9: *Summary plots for the points resulting from the clustering algorithm. Inset: (left to right) Histogram density estimation using 20 bins, and kernel density estimation.*

contrast to plot 12, which also has high variance, however, this seems to be due to a small number of outliers, rather than a strong, divergent prediction. Interestingly, the location of the strongest prediction in plot 12 is not co-located with the mean. This suggests that while plot 12 strongly predicts for one particular outcome, using the mean as an estimation of that outcome is a poor choice. Conversely, plot 11 predicts two outcomes with high ensemble votes, the largest subset of outcomes being very close to the mean, and thus the mean does seem like a good choice. This type of information helps scientists understand the origins of member uncertainty; large variation stemming from member disagreement should be treated differently than the influence of a small number of outliers. While colormaps of mean and standard deviation colormaps do give indications of where such variation exists they do not allow for the greater understanding of the underlying data that is exposed through the summary plots. Also, the box plot alone does not convey the richness of this data, as shown by comparing the summary plots and their corresponding box plots in Figure 9.



☑ Draw Box Plot

☑ Draw Density

☑ Draw Central Moments

☑ Draw Higher Order Statistics

☑ Draw Best Fit Distribution

Figure 10: *(Left, top) Close-up of the joint summary plot for multiple categorical datasets. While the joint histogram display induces some visual clutter, the cloud-like nature of the display gives a general feel of the density trend across the data. The covariance and skew variance glyphs help distinguish between each joint summary plot. (Left, bottom) A dialog box from the user interface. (Right) 2D summary plot of temperature and humidity averaged across altitude slices. The trend of both variables to condense as altitude increases is visible through the compaction of the joint density display and the reduction of the size of the covariance glyph. The changes in orientation of the skew variance glyphs highlight the dominance of the temperature variable at higher altitudes.*

Another interesting feature of the summary plots is the varying width of the histogram. This is a result of the number of bins used to calculate the histogram: the less bins used, the smoother the histogram. The number of bins chosen is important, especially when the number of samples is small. If a smoother density display is desired, kernel density estimation [Par62] can be used to estimate the underlying data distribution. The results of a 20-bin histogram compared to kernel density estimation on the same data can be seen in Figure 9, inset right. While these two plots are not meant to be a direct comparison because they show the same data at different levels of smoothing, the different visual results between the two types of density estimation are apparent. For applications in which the underlying distribution is more important than analyzing the particular data samples, kernel density estimation should be used.

Using the 2D summary plots, we further explore the data. In this case, we summarize across levels of altitude to get an understanding of how the variables change together with height. Figure 10 shows 2D summary plots for temperature and humidity at 28 levels of altitude. Each altitude slice is displayed as a 2D summary plot, and the two variables are shown as 1D summary plots along the axes for alignment. The joint density display shows the trend of both temperature and humidity as altitude increases. The higher the altitude, the less variation exists across the domain; this is visible in the smaller area taken up by the joint density display. Likewise, the covariance glyphs change from being long, stretched out ellipses, to fatter and more circular, and then small, skinny ellipses at the highest altitudes, indicating the domination of the temperature variable. The skew variance glyphs emphasize the domination of temperature through their evolution towards alignment with the temperature axis. Overall, this plot accentuates the relationship between temperature and humidity as altitude level increases in that both variables are influential at lower altitudes, but temperature is the commanding variable at higher elevations.

5.1. Discussion

The main goal of the summary plot is to create a signature for data distributions, providing for fast recognition of interesting properties. The higher-order glyphs clearly display deviations from a normal distribution and are easily compared. Because the statistical meanings of the moments are more complicated than the box plot, there will be a learning curve associated with understanding the additional information. However, due to the simplicity of the technique, a user who has learned it will easily recognize desired characteristics. The summary plot also reduces the amount of information needed to convey the data distribution–a desirable reduction when the amount of information is too large to easily understand, for example when dealing with large-scale data sets.

While our design of the summary plot attempts to create glyphs that can be presented together harmoniously and minimize visual clutter, the presentation of all of this information at once may still be overwhelming. To ease this problem, we have designed a user interface that allows the user to interactively choose the desired information to investigate, a portion of which can be seen in Figure 10, bottom inset. The user may choose to turn on or off each piece of the summary plot (i.e., density, moment plots, etc), zoom in on areas of interest, modify distribution fitting parameters, and query the plots for quantitative attributes, such as the number of samples or the value of specific statistical measures. The user interface is useful for the 1D plots however is indispensable for the 2D plots. While the cloud-like structure of the density display does visually describe the density relationship present between data sets, it can occlude regions of the plot. Thus, the interface allows the user to remove the density display at will, as well as choose to not show subsets of the data. Perspective can also be problematic in the 2D plots since the size of the variance glyphs is related to how far back they are displayed. Again, this is eased by the control of the user over the viewpoint and the ability to query specific values through dialog boxes.

The higher-order moments are very sensitive to noise, outliers, and variations in sample size. This can be problematic when the number of samples is not large enough to adequately characterize the underlying distribution. In such cases, the histogram visualization becomes extremely important. The visualization provides a redundant encoding of the characteristics expressed by the moments and also clearly shows the user that the summary is based on a sparse number of samples. Work has been done to investigate methods for calculating higher-order moments in the presence of noise, such as [GH97], however these approaches increase the complexity of calculating the moments and are often application-dependent. We have chosen to use the more simplistic formulation of moments and rely on the redundancy of the summary plot to highlight unreliabilities in the moment summary.

6. Conclusion

Uncertainty information has been inadequately addressed in the visualization community, largely because of the difficulties involved with visually expressing this additional data. If visualization is to become a robust decision-making tool, it must represent uncertainty, in some form, to the audience. This work provides a method for investigating visual characteristics of a data distribution, both for learning about the shape of the data set and for expressing the associated uncertainty.

The 1D and 2D summary plots provide a simple way to annotate features of a distribution, enhance distinguishablity between datasets, and allow for the straightforward comparison of multiple distributions. They contain, by nature, uncertainty information expressed foremost by standard deviation, but also through the higher order characteristics of the distribution. In comparison to the box plot alone, the summary plot quickly exposes salient features of the data set, such as the existence and location of outliers, the amount of variability, and the skewness of a distribution. The presentation of data in a summarized and easy to read form can quickly communicate information about large amounts of data and the data's uncertainty, emphasizing meaningful characteristics and facilitating visual comparisons.

This work is the basis for further work in uncertainty visualization, as well as the visualization of large-scale, multidimensional data. Such summarization methods are increasingly important as the size and complexity of data sets grows and visual reductions of dimensionality are required. When accompanied by higher dimensional visualization techniques, such as volume rendering or isosurfacing, summary plots are an eloquent approach for presenting drill down information, for example, when regions of interest are chosen by a user or automatically. Continuing development of the summary plot includes the examination of higher dimensional distribution data. While a direct extension of the summary plot into higher spatial dimensions may not be effective, using descriptive statistics with a visual signature to highlight notable features may prove valuable. In addition, combining information visualization and graphical data analysis methods with preexisting scientific visualization methods in a user guided setting will further facilitate the understanding of data.

7. Acknowledgements

Thanks to Samuel Gerber for providing clustering expertise and code. This work was funded in part by the DOE SciDAC Visualization and Analytics Center for Enabling Technologies (www.vacet.org), NIH NCRR Grant No. 5P41RR012553-10 and KAUST (KUS-C1-016-04).

References

[BE92] BAIN L. J., ENGELHARDT M.: *Introduction to Probability and Mathematical Statistics*. Duxbury Press, 1992.

[Ben88] BENJAMINI Y.: Opening the box of a boxplot. *The American Statistician 42*, 4 (November 1988), 257–262.

[BG87] BECKETTI S., GOULD W.: Rangefinder box plots. *The American Statistician 41*, 2 (May 1987), 149.

[Bis06] BISHOP C. M.: *Pattern Recognition and Machine Learning*. Springer, 2006.

[CC06] COHEN D. J., COHEN J.: The sectioned density plot. *The American Statistician 60*, 2 (May 2006), 167–174.

[CCKT83] CHAMBERS J. M., CLEVELAND W. S., KLEINER B., TUKEY P. A.: *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[Cle94] CLEVELAND W. S.: *The Elements of Graphing Data*. Hobart Press, 1994.

[CM05] CHOONPRADUB C., MCNEIL D.: Can the box plot be improved? *Songklanakarin Journal of Science and Technology 27*, 3 (2005), 649–657.

[DT00] DOANE D. P., TRACY R. L.: Using beam and fulcrum displays to explore data. *The American Statistician 54*, 4 (November 2000), 289–290.

[EB03] ESTY W. W., BANFIELD J. D.: The box-percentile pot. *Journal of Statistical Software 8*, 17 (2003).

[FHI89] FRIGGE M., HOAGLIN D. C., IGLEWICZ B.: Some implementations of the box plot. *The American Statistician 43*, 1 (February 1989), 50–54.

[GH97] GRUBER M., HSU K.-Y.: Moment-based image normalization with high noise-tolerance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 2 (1997), 136–139.

[GI92] GOLDBERG K. M., IGLEWICZ B.: Bivariate extensions of the boxplot. *Technometrics 34*, 3 (August 1992), 307–320.

[Hae48] HAEMER K. W.: Range-bar charts. *The American Statistician 2*, 2 (April 1948), 23.

[HN98] HINTZE J. L., NELSON R. D.: Violin plots: A box plot-density trace synergism. *The American Statistician 52*, 2 (May 1998), 181–184.

[Joh04] JOHNSON C. R.: Top scientific visualization research problems. *IEEE Computer Graphics and Applications 24*, 4 (July/August 2004), 13–17.

[JS03] JOHNSON C. R., SANDERSON A. R.: A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications 23*, 5 (2003), 6–10.

[MTL78] MCGILL R., TUKEY J. W., LARSEN W. A.: Variations of box plots. *The American Statistician 32*, 1 (February 1978), 12–16.

[NCE] NCEP: Short-range ensemble forecasting. http://wwwt.emc.ncep.noaa.gov/mmb/SREF/SREF.html.

[Par62] PARZEN E.: On estimation of a probability density function and mode. *The Annals of Mathematical Statistics 33*, 3 (1962), 1065–1076.

[Pot06] POTTER K.: Methods for presenting statistical information: The box plot. *In Hans Hagen, Andreas Kerren, and Peter Dannenmann (Eds.), Visualization of Large and Unstructured Data Sets, GI-Edition Lecture Notes in Informatics (LNI) S-4* (2006), 97–106.

[RRT99] ROUSSEEUW P. J., RUTS I., TUKEY J. W.: The bagplot: A bivariate boxplot. *The American Statistician 53*, 4 (November 1999), 382–287.

[Spe52] SPEAR M. E.: *Charting Statistics*. McGraw-Hill, 1952.

[Ton05] TONGKUMCHUM P.: Two-dimensional box plot. *Songklanakarin Journal of Science and Technology 27*, 4 (2005), 859–866.

[Tuf83] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[Tuk77] TUKEY J. W.: *Exploratory Data Analysis*. Addison-Wesley, 1977.

[Wil99a] WILKINSON L.: Dot plots. *The American Statistician 53*, 3 (August 1999), 276–281.

[Wil99b] WILKINSON L.: *The Grammar of Graphics*. Springer-Verlag New York, Inc., 1999.