A Framework for Uncertainty-Aware Visual Analytics

Carlos D. Correa * Yu-Hsuan Chan † Kwan-Liu Ma ‡

University of California at Davis

ABSTRACT

Visual analytics has become an important tool for gaining insight on large and complex collections of data. Numerous statistical tools and data transformations, such as projections, binning and clustering, have been coupled with visualization to help analysts understand data better and faster. However, data is inherently uncertain, due to error, noise or unreliable sources. When making decisions based on uncertain data, it is important to quantify and present to the analyst both the aggregated uncertainty of the results and the impact of the sources of that uncertainty. In this paper, we present a new framework to support uncertainty in the visual analytics process, through statistic methods such as uncertainty modeling, propagation and aggregation. We show that data transformations, such as regression, principal component analysis and k-means clustering, can be adapted to account for uncertainty. This framework leads to better visualizations that improve the decision-making process and help analysts gain insight on the analytic process itself.

Keywords: Uncertainty, Data Transformations, Principal Component Analysis, Model Fitting

1 INTRODUCTION

The goal of analytical reasoning is to gain insight from large amounts of disparate and conflicting data with varying levels of structure. Visual analytics seeks to facilitate this process by means of interactive visual metaphors. However, limitations on technology and human power make it difficult to cope with the growing scale and complexity of data. Therefore, it is seldom possible to analyze data in its raw form. It must be transformed to a suitable representation, which facilitates the discovery of interesting patterns. Dolfing [9] describes the visual analytics process as a series of transformations that facilitate insight from a collection of heterogeneous data sources. Thus, transformations can be categorized as *data/visual transformations*, which derive representations with increasing structure and meaning, and *visual mappings*, which convert these structures into visual elements, used by a visualization interface. Figure 1 shows an overview of such a visual reasoning process.

Data is inherently uncertain and often incomplete and contradictory. For instance, data collected from online news sources and blogs is often populated with misinformation and deception. Measured data contains errors, introduced by the acquisition process or systematically added due to computer imprecision. For the analyst, it is important to be aware of the sources and degree of uncertainty in the data. As data is pre-processed, transformed, and mapped to a visual representation, this uncertainty is compounded and propagated, making it difficult to preserve the quality of data along the reasoning process. In this paper, we present a framework to represent and quantify the uncertainty through a series of data transformations. With an explicit representation of uncertainty at all stages of the process, the analyst can make informed decisions based on the levels of confidence of the data and evaluate the insight gained on previous stages of the reasoning process. This uncertainty is not only propagated from the original data to the visual representations, but also data transformations themselves generate additional uncertainties. For example, complex multi-variate data is often projected to a low dimensional space for easy visualization, such as Principal Component Analysis, which implies a loss of information. For the user, it becomes important to have a visual representation that not only summarizes the uncertainty of the information being presented, but also helps identify the sources of that uncertainty.

To illustrate the uses of our framework, we use a case study from the Boston neighborhood housing price data set, consisting of a multi-variate dataset about different factors that affect the mean value of housing in the Boston area, collected in the 1950s [13]. This dataset is inherently uncertain, due to statistical sampling or errors. It was soon noted that a certain variable contained an incorrect bias. To analyze this data, we follow common data analysis tools, such as model fitting, principal component analysis and clustering, and show how uncertainty is not only propagated, but also aggregated in each of these stages. An uncertainty-aware framework not only reports the sources of uncertainty, but also requires transformations that can deal with uncertain inputs. We enhance traditional visual analytics tools with uncertainty information, which shows an overview of the distribution of error and probabilistic variance of the data. In addition, an explicit visual representation of the sensitivity coefficients reveals correlations between the output uncertainty and specific input variables that may be difficult to discover or easily missed by means of statistic analysis alone.

Keim et al. suggested the mantra: "Analyze First - Show the Important - Zoom, Filter and Analyze Further - Details on Demand" to guide the visual analytics process [17]. In our work, we show that a similar guide applies to uncertainty. We first analyze the data in terms of sensitivity and uncertainty, we show the important, i.e., the most influencing or uncertain variables and then we show details on demand, such as sensitivity coefficients for specific transformations and data points. In this paper, we present a series of visual representations that combines summarized and detailed views of the uncertainty of a multi-dimensional complex data set. Although a proof-of-concept case is depicted, we believe our framework can be extended to incorporate a variety of visual analysis tools.

2 RELATED WORK

Multivariate analysis is at the core of visual analytics. Methods for this type of analysis include regression [10], generalized additive models [14] and response surface analysis [3]. These methods in general try to find relationships among variables and fit models to multi-variate data. Other tools are used to reduce the amount of information encoded in the multi-variate data, such as binning and sampling [28], projections [25], multi-dimensional scaling [4] and clustering [2].

Yang et al. integrate analysis tools with visual exploration of multivariate data [33] using the Nugget Management System, which incorporates user interest to guide the analysis. Barlowe et al. introduce the derivatives of dependent variables to help find correlations between variables [1]. In our paper, we study another

*correac@cs.ucdavis.edu

†yhcchan@ucdavis.edu

‡ma@cs.ucdavis.edu

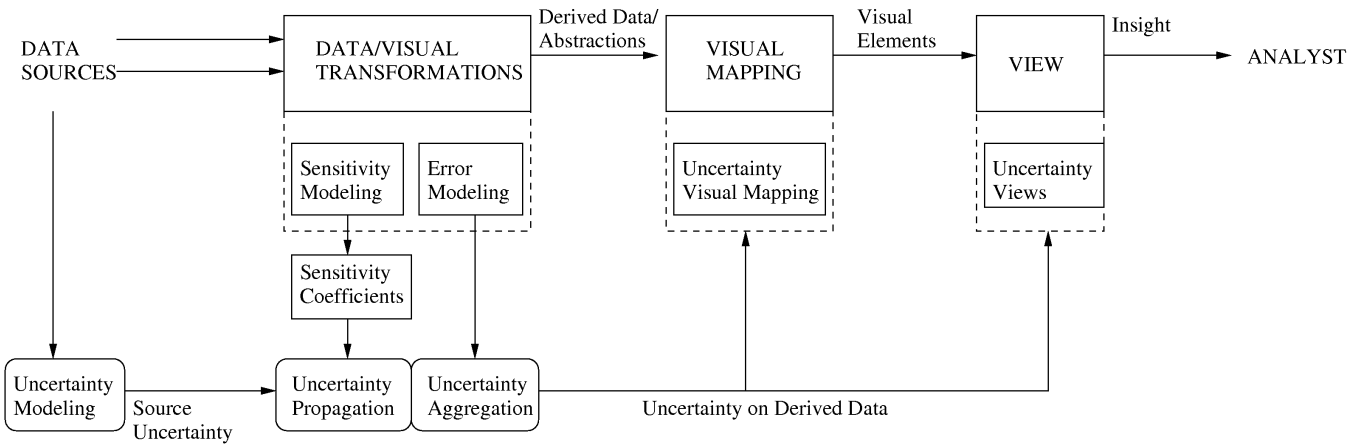


Figure 1: Uncertainty-aware Visual Analytics Process. In general, visual analytics is the process of transforming input data into insight. A similar process occurs for the uncertainty. First, *uncertainty modeling* generates a model for source uncertainty. As data is transformed, these uncertainties are propagated and aggregated. We obtain such estimates via sensitivity and error modeling. Finally, the uncertainty on the derived data and its sources are mapped to visual representations, which finally populate the view used by the analyst.

important aspect when dealing with multivariate data, the issue of data and transformation uncertainty.

Although there is no consensus on the scope of uncertainty, a number of definitions have been proposed. Hunter and Goodchild [16] define uncertainty as "the degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations without caution". This definition has spun a number of interpretations on what can be measured as uncertainty. Taylor and Kuyatt [27] proposed a series of guidelines for evaluating uncertainty of measurement results, which is classified as either random or systematic error. This led to a classification by Pang et al.[21], who suggested three types of uncertainty that are relevant for visualization of complex data: *statistical*, for measurements with a known distribution, *error*, a difference between a measure and known ground truth, and *range*, which represents an interval where data exists. To accommodate the varying data types and transformations on an analytical process, Thomson et al. define an uncertainty typology [29], identifying key components of uncertainty such as accuracy/error, precision, completeness, consistency, lineage, credibility, subjectivity, and interrelatedness. These frameworks have stemmed from geospatial information systems. Recently, Zuk and Carpendale extended this typology of uncertainty to reasoning as a way to support visualization of analytic processes [35].

The study of uncertainty can be further categorized as those concerned with uncertainty modeling and those with uncertainty propagation. To model uncertainty, numerous techniques have been proposed, including probabilistic measures, Bayesian networks [18], belief functions [12], interval sets [34] and fuzzy sets [22]. A different issue is the process of uncertainty propagation, which deals with the fact that uncertainty gets transformed as data moves through the analytics process. As suggested by Taylor and Kuyatt for the analysis of variance of measurements, it is possible to derive the variance propagated by a transformation as a linear combination of the variance of its inputs [27]. This simplification is also valid for non-linear transformations, as it results from the first order Taylor expansion of the transformation. This model was further simplified by Thomson et al. [29], who proposed to model the different uncertainty types of their typology as the output of simple operations on its variances. In their case, the uncertainty of transformations or the analysis process (often a subjective measure) is known. In general, the uncertainty of a transformation must be derived from a sensitivity or perturbation analysis, which involves the differentiation of a transformation with respect to its inputs. Because some transformations are only provided as black boxes, these parameters must be approximated, using methods such as linear least squares and expectation-maximization (EM) algorithms. Frey and Patil review a number of sensitivity analysis methods [11]. Tanaka surveys the sensitivity analysis in the scope of multivariate data analysis [26]. Specific analyses of uncertainty for certain common data analysis tools have been proposed. Chan et al. present a sensitivity analysis for variance-based methods in general [5]. Cormode et al. [7], Chau et al. [6] and Ngai et al. [20] propose extensions to perform k-means clustering on uncertain data. Similar studies have been carried out to quantify the sensitivity and uncertainty of the principal components of multi-variate data [30, 32]. Kurowicka and Cooke extend the issue of uncertainty analysis with high dimensional dependence modeling, combining both analytical tools with graphic representations [18]. In most of these cases, sensitivity analysis implies knowing the derivatives of the data transformations. Barlowe et al. incorporate derivatives to help the analyst assess the sensitivity of dependent variables with respect to the source data [1]. In our paper, we use the derivatives of data transformations in a more general way, as a means to measure and quantify uncertainty propagation and aggregation throughout the visual analytics process.

3 Uncertainty Framework

The visual analytics process is often described as a sequence of transformations from raw data to insight, via abstractions and visual representations, as depicted in Figure 1. The process of transforming raw data to abstractions and derived data is in fact a complex network of transformations, which propagates and aggregates uncertainty. To measure this uncertainty, we augment the visual analytics in the following ways:

First, the input data uncertainty is modeled. Uncertainty modeling is a rather general approach, and numerous methods have been proposed to achieve it, including parametric and non-parametric models. In the former, a statistical model is applied to the input data. In the case of Gaussian distributions, for example, the standard deviation serves as a measure of the data uncertainty. In the latter, the uncertainty is represented from the data distribution directly, e.g., as a histogram.

As the data is transformed, this uncertainty is propagated through the analytic process. The amount of uncertainty propagated by a transformation depends on how *sensitive* is the output given a set of inputs. To model the uncertainty propagation, we extend data

transformations so that we can query their *sensitivity parameters*. In addition, transformations themselves aggregate uncertainty, typically due to error or loss of information. We obtain the aggregated uncertainty via error modeling of the transformation. Transformations that can be queried for their sensitivity parameters and the aggregated error are known as *uncertainty-aware* transformatins.

Once the uncertainty is modeled and propagated, the results is propagated through the visual mapping stage, via an uncertainty visual mapping and uncertainty views. The following sections describe each of these stages in detail.

3.1 Uncertainty Modeling

Uncertainty modeling consists of deriving a mathematical model to describe the uncertainty of the source data. There are numerous methods for modeling uncertainty, including probability measures, belief functions, interval arithmetic and fuzzy sets [12]. In this paper, we focus on parametric models and consider the input variables as random variables. In this paper, we consider the input data X to be modeled as a Gaussian Mixture Model (GMM):

$$X \sim \sum_{i=0}^{N} N(\mu_i, \sigma_i) \tag{1}$$

where $N(\mu, \sigma)$ is a Gaussian distribution of mean μ and standard deviation σ. The uncertainty is then represented as a collection of standard deviations.

Gaussian mixture models are interesting since they can be adapted to fit a wide range of probability distributions. Although non-parametric models are becoming very popular for representing uncertainty, we believe that the use of parametric models is useful in applications when the analyst aims at deriving quantitative models that explain the distribution of data.

It is important to categorize the uncertainty in the visual analytics process as either *data* or *transformation uncertainty*. The former deals with the uncertainty inherent in the source and derived data, either due to error, incomplete data or source reliability. The latter represents the uncertainty added by the data transformation. In the visual analytics process, we must be able to represent the interaction between these two types of uncertainty. This is achieved through methods such as uncertainty propagation and aggregation. Figure 2 shows an example of Gaussian models used to estimate the uncertainty in the distributions of a number of variables from a housing data set [13].

3.2 Uncertainty Propagation

Uncertainty propagation occurs as data is transformed along the visual analytics process. The more sensitive a given transformation is to variation, the larger is the uncertainty propagated through it.

For the case of Gaussian distributions, it is well known that uncertainty propagates linearly for linear transformations. This generalizes to mixtures of Gaussians as well. Non-linear models, however, do not result in uncertainty propagated using the same transformation, but can be approximated as follows.

Let us consider a nonlinear transformation :

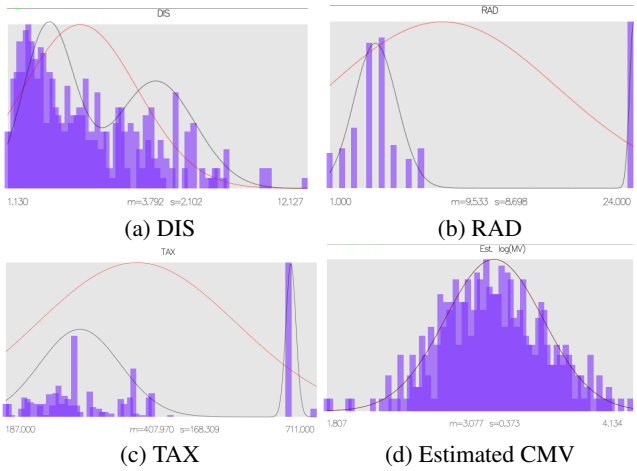
$$\mathbf{y} = f(\mathbf{x}) \tag{2}$$

of a multi-variate vector $\mathbf{x}$. Let $\mathbf{x}$ be modeled as a Gaussian mixture model:

$$p(\mathbf{x}) = \sum_{i=1}^{N} w_i N(\mathbf{x}|\mu_i, \Sigma_i) \tag{3}$$

where $N(x|\mu, \Sigma)$ is a Gaussian probability distribution with mean μ and covariance Σ.

The result of applying f is another Gaussian mixture model with mean μ' and covariance Σ', such that:



(a) DIS (b) RAD

(c) TAX (d) Estimated CMV

Figure 2: Modeling the uncertainty of variables using Gaussian Mixture Models (GMM) for 4 different variables. The first two (a-b) are accessibility variables, (c) is a neighborhood variable and (d) is a derived variable as the model fitting of the 14 variables. A Gaussian model (red) fails to capture the different peaks and misrepresents the uncertainty of the distribution. A GMM models more closely the actual distribution of the data.

$$\mu' = f(\mu) \tag{4}$$

$$\Sigma' = J(\mu)\Sigma J^T(\mu) \tag{5}$$

Where J is the Jacobian of the transformation, i.e.,

$$J_{ij} = \frac{\partial \mathbf{y}_i}{\partial \mathbf{x}_i} \tag{6}$$

This linearization of the uncertainty propagation (modeled as the covariance) derives from the first order Taylor approximation of f. Alternatives to this method include Montecarlo sampling [15], Moment methods [23] and Polynomial Chaos [31], primarily used in risk assessment and engineering uncertainty.

In the Taylor series method, used in our paper, the main issue is estimating and representing the sensitivity parameters of the transformation, described next.

3.2.1 Sensitivity Parameters

To apply the previous propagation equation, the framework requires to know the Jacobian of the transformations, formed by the partial derivatives of the transformation with respect to its inputs. These are also known as the *sensitivity coefficients* of the transformation.

There are numerous methods for finding sensitivity coefficients [11]. In this paper, we consider two of them: *analytical* and *linear regression*. To compute the sensitivity coefficients analytically, we must know the transformation in its analytic closed form, as a function in terms of the input variables. The derivatives can then be obtained symbolically and applied to the inputs.

Another alternative is to approximate the partial derivatives via linear regression. This is obtained by considering the Taylor approximation of an output variable for a number of N samples:

$$y_i = y_0 + \frac{\partial y}{\partial x}(x_i - x_0) \tag{7}$$

Using linear least squares, we obtain the approximation to the partial derivatives as:

$$\frac{\partial y}{\partial x} \approx \frac{\sum_{i=0}^{N}(y_i - y_0)(x_i - x_0)}{\sum_{i=0}^{N}(x_i - x_0)^2} \tag{8}$$

3.3 Uncertainty Aggregation

Data transformations, in general, involve certain error and often result in loss of information. For this reason, transformations themselves aggregate uncertainty to the output variables. Here, we consider the uncertainty of a transformation as Gaussian noise. Therefore, Eq.(2) can be extended as follows:

$$\mathbf{y} = f(\mathbf{x}) + e \quad (9)$$

$$e \sim N(0, E) \quad (10)$$

where e is an error term, modeled as a Gaussian distribution of zero mean and covariance E.

When we account for both propagation and aggregation of uncertainty, the result of applying a transformation results in the uncertainty:

$$\Sigma' = J(\mu)\Sigma J^T(\mu) + E \quad (11)$$

3.4 Transformation Uncertainty

To understand the implications of our framework, we show the analysis of uncertainty for two typical data analysis transformations: projection via Principal Component Analysis, and clustering.

3.4.1 Principal Component Analysis (PCA)

PCA is a projection method that re-expresses a collection of correlated variables into a smaller number of variables called principal components, which maximize the variance of the data. PCA has become an important analysis tool for visual analytics, as N-dimensional data can be projected into a lower dimensional space (typically 2D), which can be represented easily in current display technology.

To understand the effects of PCA in input uncertainty, we must perform a sensitivity analysis. Several methods have been proposed before to this purpose [26, 32]. Here, we follow a generic approach of multi-dimensional differentiation as described in the previous section.

In addition to this propagation, PCA itself adds uncertainty, seen as loss of information as several dimensions (the ones with least variance) are ignored. Let us consider the case of PCA into two dimensions for an $m \times n$ matrix X representing m observations of an n-dimensional vector. The PCA projection is a linear transformation:

$$Y = P^{(k)}X \quad (12)$$

where P^k is a linear transformation containing the first k principal components of X. A typical projection in 2D uses $k = 2$. Therefore, the error introduced by this projection can be computed as:

$$E_{PCA} = ||P^{(2)}X - P^{(n)}X|| \quad (13)$$

It can be seen that this is equivalent to

$$E_{PCA} = \frac{1}{2}\sum_{i=3}^{n} \lambda_i \quad (14)$$

where λ_i are the eigenvalues of the covariance matrix resulting of the empirical zero mean data matrix, which summarize the magnitude of the secondary components.

3.4.2 Clustering

Another commonly used transformation is clustering, which arranges data values in a large collection into separate classes that minimize the distance between points in the same class, while maximizing the distance between points in different classes. Methods for clustering include k-means, hierarchical algorithms, locality- and grid-based algorithms [9]. The k-means algorithm is a greedy algorithm that iteratively assigns data points to the cluster whose centroid is closest [19]. When the data is uncertain, however, the distance between data points cannot be determined deterministically. Instead, k-means must take into account the variation of the data points. An example is UK-means [6], which considers the expected distance to a cluster centroid instead of the actual Euclidean distance.

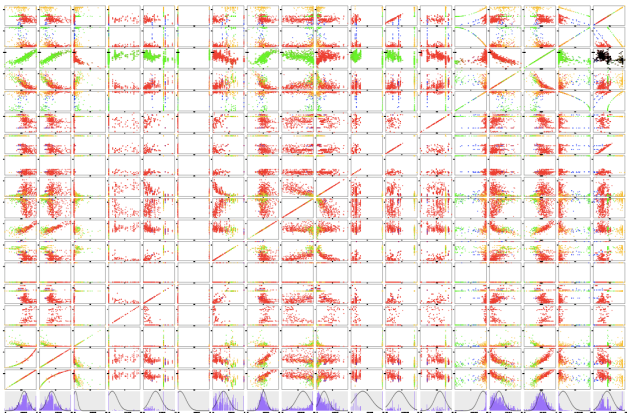


Figure 3: BNHP Dataset, represented here as a scatterplot matrix. Clearly, the high dimensionality makes it difficult to understand and find meaningful correlations.

Similar to PCA, clustering introduces error. In general, we can measure the "quality" of the clustering using the total variance that k-means is trying to minimize:

$$E = \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2 \quad (15)$$

for k clusters with centroid μ_k. x_i are the data points, classified into the sets C_k.

3.5 Visual Mapping

Finally, the uncertainty propagated and aggregated throughout the process is mapped to visual representations. Analogous to the original data, this implies a problem of visualization of multi-dimensional data, since the uncertainty of the data depends on each of the variables, the intermediate results after data transformations and the data transformations themselves. Following the visual analytics mantra, the visual mapping needs to be multi-functional. On one hand, it should provide an overview of the uncertainty, and on the other hand, it must let analysts gain access to detail information. To achieve the first, we enhance scatter plots of multidimensional data with uncertainty nodes, whose size indicate the magnitude of the uncertainty. Using transparency, we "hide" the effects of uncertainty so that only the most reliable data is highlighted to the user. A different view does the opposite: it enhances the data with higher uncertainty. This is useful for discovering the sources of uncertainty and formulate questions about their distributions. An example is detailed in the following section. For detail information about uncertainty, we explore the use of bar charts (or tornado graphs in Cooke and Noordwijk [24]), which depict the contribution of each variable and data transformation in the uncertainty of a given data point.

4 Case Study

To test our framework, we used a case study based on the Boston neighborhood housing price data set (BNHP). This data set consists of 14 variables and 506 data records of housing market data in the

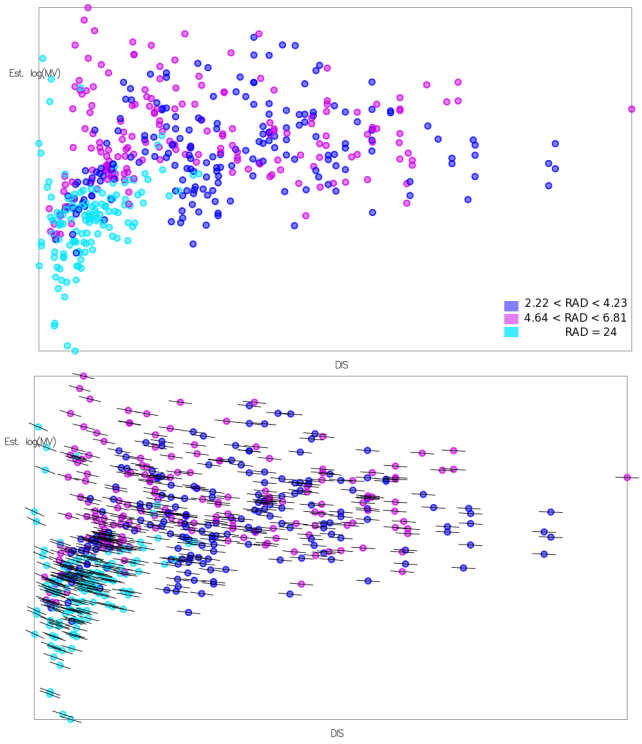


Figure 4: Derivative Visualization. To show the derivatives, we use a linear trend line for each sample value. The overall shape indicates the general direction and expected variation of a data point. Color coding denotes a clustering with respect to the variable RAD. Compare a traditional scatterplot (top) with one augmented with derivative information (bottom).

Boston metropolitan area [13]. These variables include some *structural* information such as the number of rooms in a unit (RM) or age of the building (AGE), *neighborhood* related, such as the proportion of population with lower status (LSTAT), crime rate (CRIM), proportion of nonretail business (INDUS), *accessibility* variables, such as the distance to five employment centers in Boston (DIS) and accessibility to radial highways (RAD), and an *air pollution* variable, the concentration of nitrogen oxide (NOX). The data set was collected in an effort to propose a procedural model of the willingness to pay for clean air.

4.1 Uncertainty Modeling

The BNHP dataset consists of 14 variables, which we modeled using mixtures of Gaussians. These variables are depicted in Figure 3. We can see that a scatterplot matrix like this makes it difficult to understand the complexity and correlations of multi-variate data, even with interactive capabilities such as zooming and filtering. To better depict this data set, we use principal component analysis to project the 14 dimensions in a 2D plot.

To extract the model for each variable, we estimate the probability parameters using Maximum Likelihood criterion using the Expectation-Maximization algorithm [8]. Examples of the result of GMM modeling for a number of variables in the BNHP dataset is shown in Figure 2, including a derived variable. Clearly, a simple Gaussian distribution does not capture the complex shape of their probabilistic distribution. A Gaussian Mixture Model (GMM), consisting of two Gaussian clusters, represents more accurately the uncertainty of these two variables.

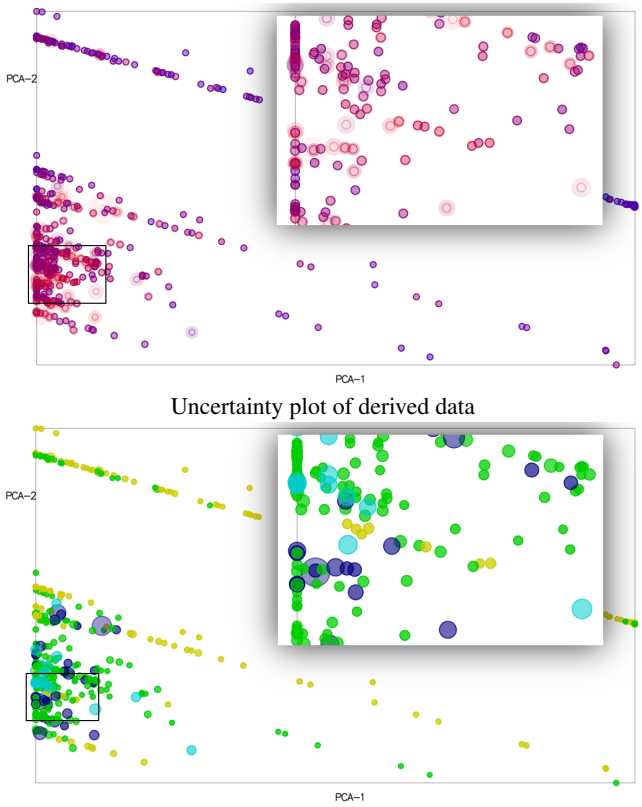


Figure 5: Uncertainty view of the housing mean value resulting from model fitting. Top: Color indicates the price, lower prices are in blue, while higher prices are in red. Size of a point denotes uncertainty. Note that the results for higher prices are consistently more uncertain than low priced housing. Also note the concentration of these nodes towards one of the ends in the PCA projection. Bottom: To understand the sources of this uncertainty, we map the most important factor that influences it (i.e., the most sensitive parameter). The color coding is defined in Figure 6. The largest uncertainty seem to be correlated with the neighborhood variables (LSTAT) while medium and low uncertainties relate to accessibility variables (RAD and DIS).

4.2 Model Fitting Uncertainty

According to Harrison and Rubinfeld [13], it is possible to define the housing value as a nonlinear combination of the 14 input variables, whose parameters can be found using nonlinear regression. Therefore, we can readily find the sensitivity parameters of this derived data with respect to each of the outputs. Examples of these are depicted in Figure 4. At the top, we show a traditional scatterplot. This plot does not tell much about the trends in the data. In the bottom, we see a scatterplot *augmented* with uncertainty information, in the form of line segments. Now the user can clearly see trends in the data. The largest the slope of these segments, the higher the sensitivity. We can clearly see a negative correlation of the estimated mean value MV with respect to the variable DIS (x-axis) and the variable RAD (clustering). Values in the cluster denoted by cyan (high RAD levels) seem to be more uncertain.

To visualize the uncertainty, we map the magnitude of the propagated uncertainty to the size of nodes in a 2D scatterplot defined as the PCA projection of the 14 variables. This is shown in Figure 5. On top, we use color to encode the value of the output variable, resulting from the model fitting of the 14 input variables. Red nodes denote high housing prices while blue nodes denote low prices. Transparency also encodes the degree of uncertainty. The

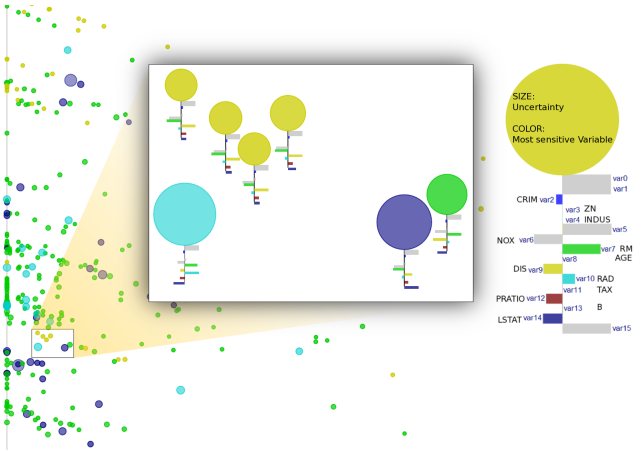


Figure 6: Detail uncertainty information. The different sensitivity parameters are shown for each data point as a bar or tornado chart [24] to represent both the magnitude and sign of the sensitivity. Bars to the left indicate a negative sensitivity, while bars to the right indicate a positive sensitivity.

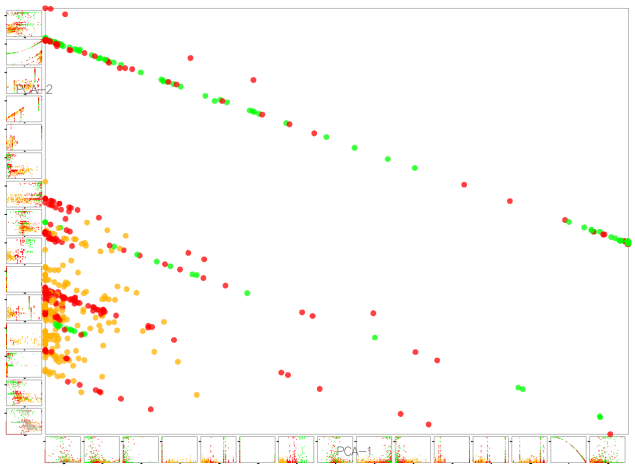


Figure 7: A summarized representation of the same data as a scatterplot of the 2 principal components. Color tagging of different clusters helps the analyst understand the dependencies between different variables. Top: Clustering based on NOX variable (air pollution). A simpler visualization that carries more information in a single view can be obtained with uncertainty, as shown in Figure 5.

more uncertain a data value is, the more transparent is its visual representation. This "hides" the effect of uncertainty and steers the user's attention towards the most reliable data points. With this visualization, we immediately see a correlation of uncertainty with the estimated median value. Highly priced housing seems to carry a lot more uncertainty than low priced housing. We also noted that they appear clustered in a particular region of the projection, suggesting a specific cause for this uncertainty.

To understand more the sources of uncertainty, we turn to the sensitivity parameters, discovered in the initial stages of our framework. The most sensitive variable turns out to be NOX (concentration of nitrogen oxides in the air), the only air pollution variable considered in the original study [13], which in turn was found to be an important variable and used to measure the willingness to pay for clean air. We then turn to the second most sensitive variable, as depicted in Figure 5-bottom. Here, instead of hiding the effects of uncertainty, we highlight them. Note how the more uncertain nodes are visible and color coded depending on the second most sensitive variable (since the most sensitive is NOX for all of them). We see a correlation of larger uncertainties with the variable LSTAT (the proportion of adults of lower status, in blue), medium uncertainties to RAD (the index of accessibility to radial highways, in cyan), while relatively low uncertainties to DIS (distance to five employment centers in the Boston region, yellow) and RM (number of rooms in owner units, in green). These suggest that more confidence can be attributed to the effect of *accessibility* variables to the mean housing price, than other *neighborhood* o *structural* variables. These results are consistent with the findings in Harrison's original study [13].

Figure 6 shows a zoomed-in view of a portion of the scatter plot depicting detail information about the sensitivity parameters. In this case, we use a tornado representation (a bar chart) [24] to show both the sign and magnitude of the partial derivatives. Again, we see a predominance of the RAD, DIS, LSTAT and NOX variables in the uncertainty of the mean housing value, but we also discover the effects of other variables. We believe that this multi-level approach for exploring uncertainty, from high level overviews (Fig. 5) to detail information (Fig 6), is essential for making reliable decisions upon the visual analysis of complex data.

4.3 PCA transformation

PCA is used to reduce the dimensionality of the data set to a 2D plot. An example is shown in Figure 7, where we plot the multidimensional data sets along the principal components and use color coding to denote three clusters based on the NOX variable (air pollution). The visualization also depicts the individual pair-wise scatterplots with each variable in the x and y axes.

To model the uncertainty propagated by the PCA transformation, we estimated the sensitivity parameters via linear regression. Because we want to observe local changes in PCA transformation according to the different input variable, we used moving least squares instead of the total least squares method in Section 3.2.1. In the moving least squares sense, the derivative of PCA with respect to a variable is computed only in a neighborhood around each data point. An example is shown in Figure 8, which depicts the sensitivity of the second principal component with respect to the input variable LSTAT. The color denotes a clustering with respect to another variable (NOX), which suggests a correlation between this sensitivity and the other variable. Notice the presence of critical points, where the sensitivities can become positive or negative. As a data point moves around this critical region, the variable changes from influencing positively the PCA projection to negative influence.

Once we have estimated the sensitivity parameters, we can estimate the uncertainty propagated by the PCA transformation. The 2D projection leads to a 2D uncertainty estimate for each data point, which can be represented as ellipses. A ellipse elongated along a given dimension, i.e., horizontal or vertical, shows a larger uncertainty due to the principal or secondary component of the projection, respectively. The uncertainty visualization is depicted in Figure 9. Notice that there is a predominance of uncertainty due to the secondary component, and that there is a spatial consistency in the uncertainty estimates. Unlike Figure 5, there is no apparent correlation between high housing prices and propagated uncertainty.

When combined with the uncertainty propagated by the model fitting, this framework helps analysts to quantify and assess their confidence of, not only the input data and models, but also the data transformations.

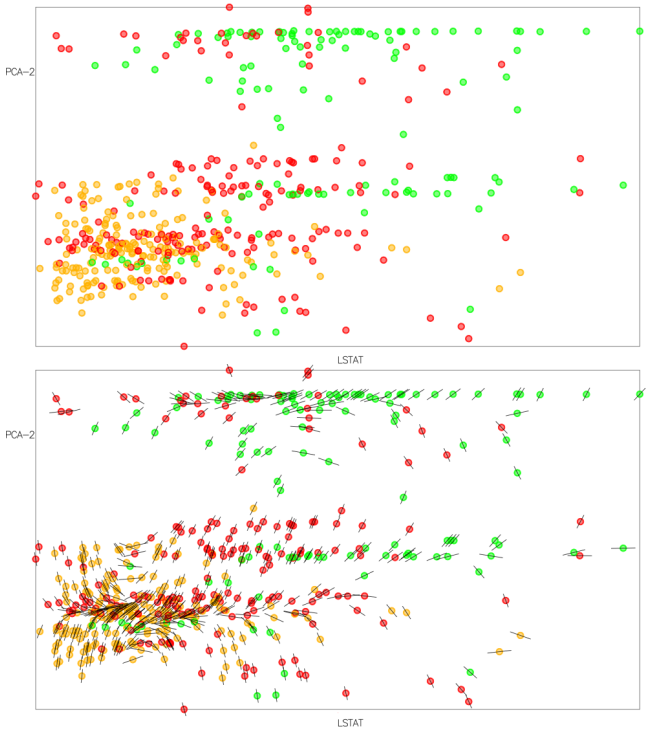


Figure 8: PCA Sensitivity to a given variable (LSTAT). Color coding denotes a clustering with respect to the variable NOX. Note that the sensitivity analysis shows us a critical region, where derivatives change sign.

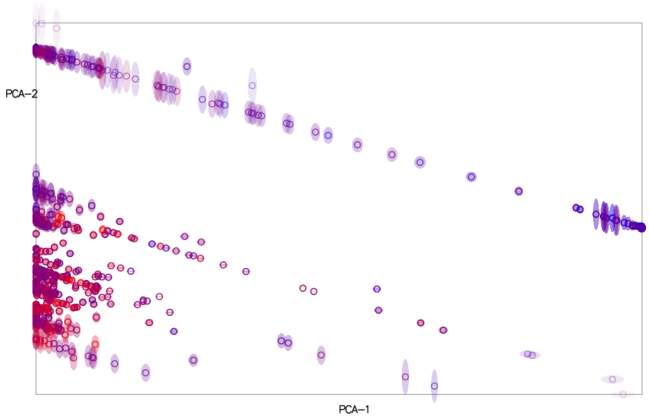


Figure 9: Uncertainty Propagation of PCA. Each component of the PCA projection propagates uncertainty, therefore, we obtain a 2D uncertainty for each data point, here represented as ellipses. The x-axis of each ellipse represents uncertainty propagated by the principal component, while the y-axis represents the uncertainty propagated by the secondary component.

4.4 Clustering

To understand the effects of clustering in uncertainty, we measure the variance within each cluster for a number of possibilities. Because clustering is an operation that maps from a large set of data points to a small number of classes, uncertainty is only represented as a summary view. Fig. 10 shows the stacked histogram of the uncertainty for clustering along different dimensions. The larger the bar, the highest the uncertainty. The stacked histogram also shows the relative uncertainty of each of the clusters. When clusters are highly uncertain in relation to others, such as for the CMV and CRIM variables, this suggests that the number of clusters is not necessarily optimal, and the analyst can improve the classification. In other cases, such as for RAD and TAX, the clustering seems to classify the data well. Fig. 10 also shows the stacked histogram for 4 clusters. We see that this change improves the uncertainty of clustering for some of the variables, such as CMV, CRIM and PRATIO. In general, the optimal number of clusters for each variable can be found using expectation-maximization techniques (EM). In other cases, however, prior information about the data may suggest some expected number of clusters that may not be optimal. Our framework enables the analyst to *evaluate* the quality of the data transformations. For example, we also show the effect of applying a more appropriate clustering to the data. In Figure 10, we show the uncertainty for variables EMV (result of model fitting) and UEMV, which represents the clustering of EMV using UK-Means [20]. We can see that, for both 3 and 4 clusters, UK-Means generates a clustering with less variance than the counterpart that does not include uncertainty. With our framework, analysts are able to compare quantitatively the efficacy of their data transformations.

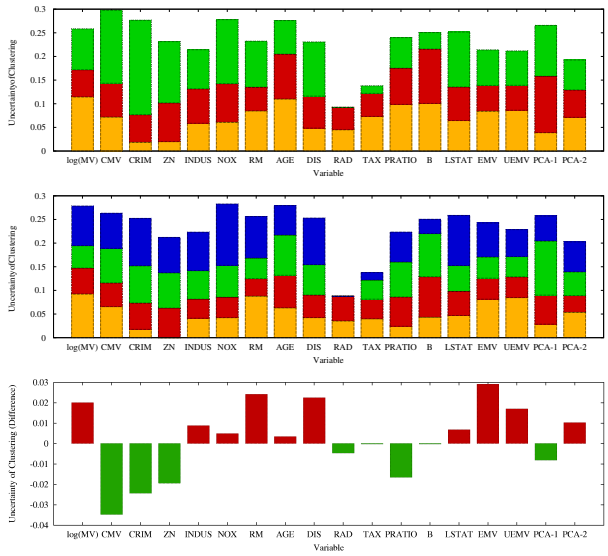


Figure 10: Uncertainty of Clustering. Each bar represents the total variance of the clustering operation for each variable, including the two principal components of the data. Within each bar, the size of each color represents the uncertainty of each cluster. The first and second rows show the uncertainty for 3 and 4 clusters, respectively. The last row shows the difference in uncertainty. A negative difference (green) indicates an improvement of uncertainty. This summarized view helps the analyst evaluate the efficacy of the data transformations.

5 Limitations and Conclusion

We have presented a general framework for introducing uncertainty in the visual analytics process. We found that mirroring the process of transforming data into insight allows us to define a series of operations on uncertainty, such as modeling, propagation and aggregation, that map input uncertainty to visual representations.

We have followed a quantitative approach that models uncertainty as the propagation and aggregation of error in a parametric model of the distribution of data. We believe that this mechanism is useful when the analyst wants to extract a model that explains the behavior of data and helps make projections or extrapolate to different situations. Discrete operations, such as clustering, can also be included in the framework, provided a quantitative measure of

uncertainty. More qualitative assessment of the uncertainty is difficult to model in our framework. We believe that our framework can be extended with Bayesian networks to support more general data types. Another aspect of our approach is its scalability. Sensitivity analysis is at the core of the framework, which requires analyzing the effects of every output variable with respect to its inputs. For extreme large data sets, this may be prohibitive. Two solutions to this problem are: (1) either provide a simplification of the data distribution and estimate the sensitivity coefficients with respect to the simplified distribution, or (2) perform uncertainty analysis locally. The former approach is useful for overviews and the latter for detailed views of the uncertainty.

The study of uncertainty proves important for understanding the sensitivity of the output with respect to the inputs. On one hand, uncertainty provides a summarized quantity for each data point, which helps the analyst assess the confidence level on the visual representation. For example, overviews of the uncertainty helped us determine a correlation with certain clusters with the confidence level. On the other hand, output uncertainty is a complex multidimensional dataset, which can be further inquired to gain access to detail sensitivity information. We applied common sensitivity visualization tools such as tornado maps and color encodings to show the correlation between uncertainty and specific variables in a multi-dimensional data set. We believe that a similar mapping can be obtained to other common visualizations, such as parallel coordinates and radar views. With the use of general methods such as Gaussian Mixture Models, statistical linearization of sensitivity parameters and uncertainty propagation, we have a framework that can be adapted to a wide variety of probability distributions and data transformations. Although the case study shown in this paper focuses on model fitting and principal component analysis, our approach can be followed to extend other data transformations, such as binning, multidimensional scaling and self-organizing maps, to account for their uncertainty.

ACKNOWLEDGEMENTS

This research was supported in part by the U.S. National Science Foundation through grants CCF-0938114, CCF-0808896 and CCF-0811422 and by Hewlett-Packard Laboratories. We also thank Nokia Research Center for their support.

REFERENCES

[1] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. pages 147–154, Oct. 2008.

[2] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[3] G. Box and N. Draper. *Empirical Model-Building and Response Surfaces*. John Wiley & Sons, 1987.

[4] J. Carroll and P. Arabie. Multidimensional scaling. *Annual Review of Psychology*, 31:607–649, 1980.

[5] K. Chan, A. Saltelli, and S. Tarantola. Sensitivity analysis of model output: variance-based methods make the difference. In *WSC '97: Proceedings of the 29th conference on Winter simulation*, pages 261–268, 1997.

[6] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In *Proc. of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006)*, pages 199–204, 2006.

[7] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 191–200, New York, NY, USA, 2008. ACM.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[9] H. Dolfing. A visual analytics framework for feature and classifier engineering. Master's thesis, University of Konstanz, 2007.

[10] N. R. Draper and H. Smith. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. John Wiley & Sons Inc, 2 sub edition, 1998.

[11] H. Frey and S. Patil. Identification and review of sensitivity analysis methods. *Risk Analysis*, 22(3):553–578, 2002.

[12] J. Y. Halpern. *Reasoning about Uncertainty*. The MIT Press, October 2003.

[13] D. J. Harrison and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978.

[14] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[15] F. O. Hoffman and J. S. Hammonds. Propagation of uncertainty in risk assessments: The need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis*, 14(5):707–712, 1994.

[16] G. Hunter and M. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. *Journal of Urban and Regional Information Systems Association*, 5(2):55 —62, 1993.

[17] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *IV '06: Proceedings of the conference on Information Visualization*, pages 9–16, 2006.

[18] D. Kurowicka and R. Cooke. *Uncertainty Analysis with High Dimensional Dependence Modeling*. Wiley, 2006.

[19] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[20] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Yip. Efficient clustering of uncertain data. pages 436–445, Dec. 2006.

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

[22] B. Pham and R. Brown. Visualisation of fuzzy systems: requirements, techniques and framework. *Future Gener. Comput. Syst.*, 21(7):1199–1212, 2005.

[23] M. M. Putko, P. A. Newman, A. C. T. Iii, and L. L. Green. Approach for uncertainty propagation and robust design in cfd using sensitivity derivatives. In *AIAA 15th Computational Fluid Dynamics Conference*, pages 2001–2528, 2001.

[24] C. R.M. and V. N. J.M. Generalized graphical methods for uncertainty and sensitivity analysis. *Bashkir Ecological Journal, (Special Issue)*, 1(8):54–57, 2000.

[25] J. Shlens. A tutorial on principal component analysis, December 2005.

[26] Y. Tanaka. Recent advance in sensitivity analysis in multivariate statistical methods. *Journal of the Japanese Society of Computational Statistics*, 7(1):1–25, 1994.

[27] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, NIST Tecnical Note 1297, 1994.

[28] S. Thompson. *Sampling*. 1992.

[29] J. Thomson, E. Hetzler, A. Maceachren, M. Gahegan, and M. Pavel. A typology for visualizing uncertainty. In R. F. Erbacher, J. C. Roberts, M. T. Gröhn, and K. Börner, editors, *Visualization and Data Analysis 2005. Proceedings of the SPIE, Volume 5669*, pages 146–157, March 2005.

[30] V. Šmídl and A. Quinn. On bayesian principal component analysis. *Comput. Stat. Data Anal.*, 51(9):4101–4123, 2007.

[31] N. Wiener. The homogeneous chaos. *American Journal of Mathematics*, 60(4):897–936, 1938.

[32] Y. Yamanishi and Y. Tanaka. Sensitivity analysis in functional principal component analysis. *Computational Statistics*, 20(2):311–326, 2005.

[33] D. Yang, E. A. Rundensteiner, and M. O. Ward. Analysis guided visual exploration of multivariate data. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 83–90, 2007.

[34] Y. Yao. Interval based uncertain reasoning. *Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American*, pages 363–367, 2000.

[35] T. Zuk and M. S. T. Carpendale. Visualization of uncertainty and reasoning. In *Smart Graphics*, pages 164–177, 2007.