

# Exploring the Sensitivity of Choropleths under Attribute Uncertainty

Zhaosong Huang, Yafeng Lu, Elizabeth A. Mack, Wei Chen, and Ross Maciejewski, *Senior Member, IEEE*

**Abstract**—The choropleth map is an essential tool for spatial data analysis. However, the underlying attribute values of a spatial unit greatly influence the statistical analyses and map classification procedures when generating a choropleth map. If the attribute values incorporate a range of uncertainty, a critical task is determining how much the uncertainty impacts both the map visualization and the statistical analysis. In this paper, we present a visual analytics system that enhances our understanding of the impact of attribute uncertainty on data visualization and statistical analyses of these data. Our system consists of a parallel coordinates-based uncertainty specification view, an impact river and impact matrix visualization for region-based and simulation-based analysis, and a dual-choropleth map and t-SNE plot for visualizing the changes in classification and spatial autocorrelation over the range of uncertainty in the attribute values. We demonstrate our system through three use cases illustrating the impact of attribute uncertainty in geographic analysis.

**Index Terms**—geospatial analysis, uncertainty, visualization, choropleth.

## 1 INTRODUCTION

THE choropleth map is one of the most common methods of visualizing spatially referenced data. In creating these maps, data analysts must choose a classification scheme for binning the attribute values of data. This is an important aspect of the analysis because the choice of classification scheme directly influences the visual design. This is a relatively simple task when analyzing spatial data but becomes much more complicated in the presence of uncertain attribute values (e.g., [50], [72]). The issue of uncertainty becomes even more complex in analyses implementing multivariate classification (e.g., geodemographic profiling [6], [70]). Unfortunately, choropleth maps have a limited design space with respect to the number of visual variables that can be used to convey information. This has led to a variety of novel visualization designs for representing uncertainty in geographic data [38], [39]. Previous work has explored glyphs [63], animation [21], [25], and linked views [71] as a means of displaying uncertainty.

Given the need for better classification schemes that incorporate uncertainty in spatial data, a growing body of work has developed a series of techniques for dealing with this issue [10], [39], [46], [52], [61]. A large proportion of this work is focused on improving classification algorithms for choropleth maps in the presence of attribute uncertainty. For example, Zhang and Maciejewski [85] developed a scheme for quantifying the visual impact of classification boundary selection in choropleth maps.

Using measures of global spatial autocorrelation, they explored the visual impacts that would result in uncertainty due to the classification method used (i.e., model uncertainty). However, this work did not explore the impact of attribute uncertainty.

Despite the amount of attention dedicated to evaluating and visualizing attribute uncertainty in spatial data, two important aspects remain underexplored. The first is the movement of observations between class groups given varied levels of uncertainty. This is critical to understanding visual changes in classification results. Second, uncertainty in a multivariate context is underexplored. The majority of studies evaluate the impact of uncertainty on classification results in a univariate context [41], [42], [83]. To address these gaps, we propose a visual analytics methodology to help analysts and designers explore how the classification of unit  $i$  will change over its range of potential values. This is a critical step in both map design and spatial analysis. For map design, if labels shift, the visual appearance of the map can change (whether intentionally or not) [57]. Thus, the map designer needs to fully understand the impact that attribute uncertainty may have on their design scheme. Similarly, analysts need to understand the impact of uncertainty when exploring spatial autocorrelation and developing hypotheses about the underlying relationships between variables. This is particularly important given noted uncertainties in data sources from social media and wearable devices, as well as the well-noted sampling error with Census data from the American Community Survey (ACS) [50], [72].

We have developed a visual analytics system to assist designers and analysts in identifying the observations and attribute values that result in the largest visual or statistical changes in choropleth map design and analysis in the presence of attribute uncertainty. Our system contains three visualization components. (1) The attribute uncertainty specification view allows analysts to specify the range of attribute uncertainty they wish to analyze. (2) A simulation-based method is applied to generate all possible map classifications across the range of uncertainty for a given attribute and a fixed classification method. The associated visualization uses a novel impact river view for single attribute analysis and an impact matrix view for multivariate analysis. (3) A dual-map view and a scatter

- Z. Huang is with The State Key Lab of CAD & CG, Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: zhaosong\_huang@zju.edu.cn.
- Y. Lu and R. Maciejewski are with The School of Computing, Informatics & Decision Systems Engineering, Arizona State University, Tempe, AZ 85287. E-mail: {lyafeng, rmaciej}@asu.edu
- E. Mack is with the Department of Geography at Michigan State University, East Lansing, MI, 48824. E-mail: emack@msu.edu.
- W. Chen is with The State Key Lab of CAD & CG, Zhejiang University, Hangzhou, Zhejiang 310058 and also with the Joint Institute of Frontier Technologies, Alibaba-Zhejiang University, Hangzhou, Zhejiang 310058, China. E-mail: chenwei@cad.zju.edu.cn.

Manuscript received XXX; revised August XXX.

plot illustrate changes in classification and spatial autocorrelation. Our main contributions include:

- A visual representation of regional classification stability with respect to the underlying data uncertainty;
- A visual analytics methodology for exploring the impact of attribute uncertainty which integrates the classification results and spatial autocorrelation, and;
- A simulation-based comparison between classifications across the range of attribute uncertainty.

By iteratively classifying the map over the ranges of attribute uncertainty, we are able to identify spatial units where the visual appearance and spatial clusters of similar observations may be an artifact of data uncertainty instead of meaningful patterns and clusters. We demonstrate our proposed methodology via case studies on geospatial crime data and Census data from the American Community Survey.

## 2 RELATED WORK

Our work focuses on quantifying the effect of data uncertainty in spatial analysis and classification. In this section, we review related work in classification and geospatial uncertainty and visualization.

### 2.1 Map Classification

The goal of map classification is to group spatial units with similar underlying attribute measures. This is challenging as it involves maintaining within class homogeneity while maximizing class differences. From this perspective, while more classes may be preferred to fewer classes, the number of classes is limited by humans' ability to interpret subtle changes in color schemes as well as color limitations in commercial mapping packages [37].

To grapple with these challenges, various classification methods have been proposed for visualizing data. These techniques range from unclassified choropleth maps [77] to classic univariate methods (i.e. quantile, equal interval, and standard deviation methods [49]) of classification. Over time, more complex univariate classification algorithms [5], [17], [33] have been developed. For example, recent work in this area has focused on classification techniques that account for spatial autocorrelation between observations [78]. Multivariate map classification, on the other hand, typically involves classification over several variables using various data mining techniques, and there are hundreds of clustering algorithms (e.g., ROCK [28], ST-DBSCAN [8], etc.) that focus on summarizing large-scale multi-dimensional datasets [7].

For geospatial clustering, k-means [29] is one of the most popular clustering methods. Polczynski et al. [62] used k-means to classify geospatial data on a choropleth map with multiple feature attributes and made a comparison to other common classification methods. MobilityGraphs [81] also used k-means to reveal the movement patterns that were occluded in a flow map. Although clustering algorithms are often used to classify spatial units to generate choropleth maps, we use "classification" instead of "clustering" throughout the paper to refer to the process and the result conveyed by a choropleth.

We use k-means as our default clustering and classification method. However, it can be easily replaced by other classification methods (e.g., Jenks natural breaks optimization algorithm [33] and quantile classification [11] for single attribute classification). Although the algorithm itself could have some uncertainty in its output, we focus only on attribute uncertainty in this work.

While classification is the basis for map design, a variety of visual analytics systems have been developed to further enhance the exploration of geospatial data. For example, Koua et al. [43] proposed a self-organizing map method to visualize clusters and explore geospatial data using a unified distance matrix, projections, and component planes. Andrienko et al. [2] provided analysts with an interactive interface for guiding the process of trajectory clustering in a large dataset. Zhou et al. [86] enabled analysts to reconstruct subspaces to preserve interesting information. Clustervision [44] employed a scatter plot for supervised cluster visualization and comparison. However, these systems do not focus on how classification and clustering results might vary under uncertain input conditions. Our visual analytics system supports the overview, comparison, and exploration of map design when the underlying attributes have some degree of uncertainty.

### 2.2 Uncertainty Analysis and Visualization

Given the problems and prospects of uncertainty in traditional government data and large, unstructured big datasets, several studies have endeavored to analyze the amount of uncertainty, and the best strategies for incorporating this uncertainty in analytical results. These studies build on a history of research proposing strategies for exploring reliability and uncertainty in spatial data [46], [51], [52], [61]. For spatial data, we focus on two main types of uncertainty, positional uncertainty and attribute uncertainty [30], [69].

Positional uncertainty is with respect to the geographic information associated with an observation. Tucci and Giordano [79] define positional accuracy as "the difference between the recorded location of a feature in a spatial database or in a map, and its actual location on the ground or its location on a source of known higher accuracy." While important, work on classification and visualization of uncertainty has focused on attribute uncertainty. In this respect, measurement and sampling error are drivers of uncertainty [42]. Attribute uncertainty may also arise from varying the arrangement or spatial aggregation of data which is also known as the modifiable areal unit problem (MAUP) [24]. Unlike attribute uncertainty, which can report information about errors associated with the data, the MAUP remains unresolved [53], and its impact on statistical results unpredictable [22], [76]. That said, researchers have begun to design diagnostic techniques for selecting the most appropriate scale of analysis given the issue of MAUP [59].

The growing body of research on uncertainty may be grouped into three categories: studies dedicated to visualizing uncertainty, studies dedicated to incorporating uncertainty into classification algorithms, and visual analytical techniques for interactively analyzing uncertainty. In terms of assessing and visualizing uncertainty, Xiao et al. [83] developed an algorithm for probabilistically evaluating and visualizing the reliability of choropleth maps. Sun and Wong [74] proposed several strategies for visualizing these errors. One proposed strategy is the use of side-by-side maps, one for the point estimates, and one for the coefficient of variation to help visualize locations with the largest margins of error. A second strategy is to incorporate uncertainty in classification algorithms. Sun and Wong [74] suggested a modified natural breaks classification algorithm. A third option is to use interactive visualization tools. For example, Cliburn et al. [14] proposed a visual analytics system using coupled maps and confidence interval plots. Sun, Wong, and Kronenfeld [75] designed a map classification algorithm using the class separability criterion to determine class breaks based on statistically significant

differences between classes. A drawback of this approach is that it is susceptible to within-class heterogeneity [82].

Efforts at improving classification algorithms are ongoing. Chun et al. [42] used a Bhattacharyya distance algorithm for incorporating uncertainty in choropleth maps based on the class separability measure designed by Sun, Wong, and Kronenfeld [75]. To address the issue of within-class heterogeneity in the class separability approach for determining class breaks, Wei and Grubestic [82] used a Bicriterion Median Clustering Problem-Uncertainty (BMCP-U) approach to design a choropleth mapping scheme that incorporates estimates of uncertainty while resolving the issue of intra-class heterogeneity noted in previous studies [75]. The strength of the BMCP-U is that it does the best job of minimizing within-class variation while also incorporating spatial autocorrelation between observations.

Given the well-known issues with uncertainty in geographic data, a variety of methods have been proposed for uncertainty visualization [10]. Kinkeldey et al. [39] reviewed and discussed a series of studies for communicating uncertainty including uncertainty categories, visualization techniques, application domains, participants, and tasks. Using various color schemes [1], [58], [66] or transparency [1], [80] to encode different uncertainty levels is a common way to compare uncertainty. It has been found that using transparency is more effective than using color, texture or saturation in uncertainty representation [20], [46]. Gschwandtner et al. [27] discussed six different visual encodings, which combined different shapes, colors, and transparencies, for visualizing uncertainty. Correll et al. [15] proposed VSUPs which encodes low uncertainty data with a larger range of visual channel and high uncertainty data with smaller range to promote easy comparisons between data and attract more attention during decision making when uncertainty is high. Besides visualizing uncertainty by visual encoding, glyph-based [68] and grid-based (e.g., noise annotation lines [40] and trustree [36]) techniques have proven to be effective in analyzing and comparing uncertainty of geospatial data. Newman and Lee [60] used cylinders, cones, and multi-point glyphs to visualize the uncertainty in a volume visualization scenario.

Kinkeldey et al. [38] also reviewed studies that focus on the effectiveness of uncertainty visualization to help users make decisions. The decision tasks [38] include selection [12], [80], ranking [19], game-like tasks [9], [16], and real-world tasks [67]. Deitrick and Edsall [19] evaluated the influence of data uncertainty on decision making and users' confidence when using different uncertainty visualization methods. Deitrick et al. [18] presented an implicit approach to visualize the impact that the uncertainty of climate change will have on policy outcomes in a water model and urban area. Viard et al. [80] also found the differences between decision results with and without data uncertainty. Slingsby et al. [71] designed an interactive visualization strategy for identifying the attributes of data which drive classification results. They demonstrated the utility of this strategy with geodemographic data and discussed the issue of classifying uncertainty. A more recent innovation in visual analytics is the creation of a tool that uses a multi-criteria approach for evaluating classification schemes that considers several criteria: the number of classes, separability, unevenness, and intra-class variability [73]. The visual analytics system outlined in this paper differs from previous work in that it evaluates the movement of observations between classes based on prespecified levels of uncertainty.

### 3 THE IMPACT OF DATA UNCERTAINTY ON MAP DESIGN AND ANALYSIS

In order to explore the impact of attribute uncertainty on a chosen classification scheme, we propose two metrics for analyzing both the visual and statistical changes that might occur in a choropleth map as classification and analyses are performed across a range of uncertainty. Specifically, we focus on how many spatial units have changed their class (i.e., color), and how much spatial autocorrelation changes under the range of uncertainty.

For each spatial unit, we have a set of attributes (e.g., population, income, etc.) which will be used in the map classification. If we define an attribute value  $a$  at spatial unit  $i$  with uncertainty  $\epsilon$ , the range of values associated with  $i$  can be defined as  $\{a_i - \epsilon, a_i + \epsilon\}$ . Traditionally, classification is done simply using the value  $a_i$ . We propose to use this classification scheme as our base map. From there, for each spatial unit, we will modify the value  $a_i$  over the range of  $\{a_i - \epsilon, a_i + \epsilon\}$  at a given step size  $j$  and simulate a new classification scheme and evaluate local indicators of spatial autocorrelation. Thus, if  $\epsilon = 5$  and  $j = 1$ , for spatial unit  $i$ , we would have 11 different classifications from  $a_i - 5$ ,  $a_i - 4$ , ...,  $a_i + 5$  where the classification of  $a_i + 0$  is considered to be the base. We then compare each new classification scheme to the original classification and determine which spatial units (if any) would see a change in class label. We then use a measure of information variation to model the distribution of label changes and also evaluate the local Moran's I for each simulation.

#### 3.1 Impact on Map Classification

The impact of uncertainty on map classification results are quantified with two statistics, the count of flipped spatial units and the variation in information (VI) statistic. These two measures provide an indication of the number of changes in class labels and the difference in class labels arising from uncertainty respectively.

**Number of Flipped Units:** Given a classification method (e.g., equal interval, natural breaks, k-means, etc.), each spatial unit  $i$ , will be assigned a class label based on the attribute value (or vector of values for multivariate classification). As previously stated, for each spatial unit, we will simulate a classification across the given range of uncertainty at a step size of  $j$ . Thus, if we have a map with 10 spatial units,  $\epsilon = 5$  and  $j = 1$ , we will simulate 110 different classifications as compared to the initial classifier. In each of these simulations, we calculate the number of spatial units that changed their class as one measure of the impact of attribute uncertainty. This measure is one mechanism for assessing the visual change on the choropleth map. The number of flipped units is calculated as:

$$N(R_A, R_B) = \sum_{u \in S} \mathbb{I}(R_A[u] \neq R_B[u]) \quad (1)$$

where  $R_A$  denotes the original classification and  $R_B$  denotes a simulated result.  $S$  is the set of all spatial units, and  $R_X[u]$  represents the label of spatial unit  $u$  in  $R_X$ .  $\mathbb{I}(\cdot)$  is an indicator function.

The critical problem here is to ensure label stability by matching pairs of class labels between  $R_A$  and  $R_B$ . To do this, the class labels should be designed to minimize the number of class flips. Hu et al. [31] modeled a similar task as a maximum weighted matching (MWM) problem of a bipartite graph, and Zhang et al. [84] extended this pair matching to clustering results with different numbers of clusters. Following Hu et al.'s problem formulation, we used the Kuhn-Munkres (KM) algorithm [48] to match the

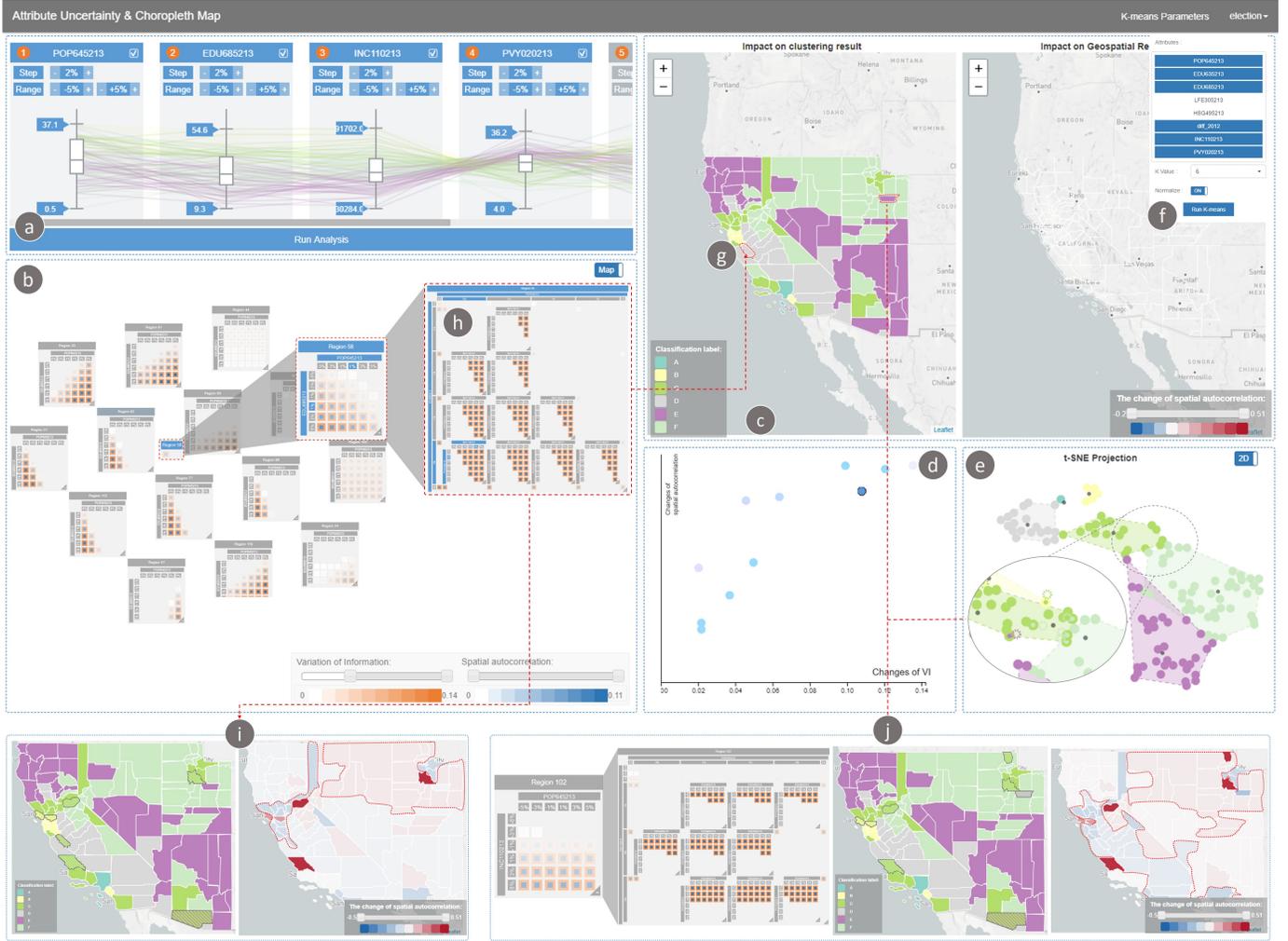


Fig. 1. (a) The attribute uncertainty specification view. (b) Region-based impact profile view. (c) The dual-choropleth view. (d) The impact quantity scatter plot. (e) t-SNE clustering Projection View. (f) The configuration panel. (g) San Benito county (h) The impact matrix of San Benito county (i) The impact of attribute uncertainty in San Benito county. (j) The impact matrix of Carbon county.

classification labels across the simulations such that the number of flips would be minimized.

To find the best matching of  $R_A$  and  $R_B$ , we generate a complete weighted bipartite graph,  $G$ , whose node set is  $\{1, 2, 3, \dots, N\} \times \{1, 2, 3, \dots, N\}$ , where  $N$  is the number of classes. The edge weights of  $G$  are calculated as:

$$W_{i,j} = \sum_{u \in S} \mathbb{I}(R_A[u] = i, R_B[u] = j) \quad (2)$$

where  $i, j \in \{1, 2, \dots, N\}$  refer to the classes. After running the KM algorithm,  $N(R_A, R_B)$  can be easily calculated. In our visualizations, we assign the matched group the same color for visual consistency (i.e., we maintain the original label from the  $a_i$  classification on as many spatial units as possible).

**Variation of Information:** While the change in class labels impacts the visual design, we are also interested in quantifying how labels change in the ranges of data uncertainty. If we consider all spatial units that belong to the same class as a cluster, then we can compare how clusters form as their attribute values change. Cluster comparison methods include counting pairs (e.g, Fowlkes-Mallows Index [23], the Rand Index [65]), set matching [45], [56], and entropy and information theoretic metrics [55]. In the family of

information theory, variation of information (VI) has been used as a metric for comparing clustering results [54]. It has been shown that VI considers not only the number of flipped units but also the resultant clustering patterns. We use the normalized term of VI to quantify the change between two classifications. We define variation of information as:

$$VI = \frac{1}{2 \log K} VI(C, C') \quad (3)$$

where  $K$  is the number of class labels,  $C$  denotes the original classification, and  $C'$  denotes the simulated classification. Specifically,  $VI(C, C')$  is the variation of information and quantifies the information differences between how data is classified in  $C$  and  $C'$ .

### 3.2 Impact on Spatial Autocorrelation

Although the changes in class labels can be measured using the approaches discussed in Section 3.1, these techniques do not provide information as to whether data uncertainty changes relationships between observations in spatial data analyses. Measures of spatial autocorrelation are often critical for identifying statistically significant local clusters, and we want to explore the stability of these measures in the presence of data uncertainty.

Spatial autocorrelation refers to spatial relationships between observations which may be positive or negative [3]. Positive spatial autocorrelation means that nearby observations have similar attribute values. Negative spatial autocorrelation means that nearby observations have dissimilar attribute values. Local spatial autocorrelation has various indicators, e.g., local G [26], Geary’s C [32], and local Moran’s I [4], [26]. We use the local Moran’s I:

$$I_i = \frac{N(x_i - \bar{x})^T \sum_j w_{ij}(x_j - \bar{x})}{\sum_j (x_j - \bar{x})^T (x_j - \bar{x})} \quad (4)$$

We define the value of each spatial unit as a vector of class indicators,  $x_i = (c_1, c_2, \dots, c_k)$ , where  $k$  is the number of classes and  $c_i$  is a binary indicator representing the class membership of unit  $i$ .  $w_{ij}$  is defined as the shared boundary between units  $i$  and  $j$ .

$$w_{ij} = \frac{l_{ij}}{l_i} = \frac{l_{ij}}{\sum_{k \neq i} l_{ik}} \quad (5)$$

where  $l_{ij}$  is the length of the boundary between units  $i$  and  $j$ .

We use the change in the local Moran’s I to quantify the impact of uncertainty on spatial autocorrelation. This impact is calculated both locally and globally in our system. We use  $I_i$  to denote the local Moran’s I for unit  $i$  before the attribute value is adjusted in our simulation process, and  $I'_i$  as the simulated value. The impact on local spatial autocorrelation is then calculated as:

$$Impact_s(i) = I'_i - I_i \quad (6)$$

The impact on global spatial autocorrelation is measured by a Euclidean distance of the local impact for all spatial units:

$$Impact_s = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^N (Impact_s(i))^2} \quad (7)$$

We are then able to explore the sensitivity of the map design based on the amount of uncertainty in attribute values.

Here, it is important to note that the calculation of measures of spatial autocorrelation are known to be sensitive to the choice of the weighting function. While our implementation utilizes the shared boundary, choices of the length of the shared boundary, Euclidean distance, etc., would result in different calculations of autocorrelation. Our methodology is appropriate for any variation in the calculation as long as the chosen weighting function remains consistent. Future work could explore how to incorporate algorithmic uncertainties into the visualization design.

## 4 VISUAL ANALYTICS ENVIRONMENT

To support spatial analysis and map design under attribute uncertainty, we have developed a visual analytics environment that consists of three views. The first view is for the evaluation of attribute uncertainty and allows users to examine attribute values and define a range of uncertainty within the data. The second view helps users quantify and summarize region-specific changes in attribute values with varying levels of uncertainty. The third view is a dual-choropleth map that helps users view and analyze changes in classification results.

### 4.1 Attribute Uncertainty Specification

For any given dataset, our system is designed to load the data and the uncertainty range by reading files, or is configurable for the analyst to assign the uncertainty range for exploration on the interface. Specifically, the uncertainty range is the lowest to

highest possible value for the attribute for the geographic unit. This means that each unit can have its own unique uncertainty range, which allows for a more robust analysis. For classification, we implemented k-means; however, this can be generalized to any classification method. For example, the system can switch to Jenks natural breaks optimization algorithm [33] and quantile classification [11] for single attribute classification. Classification parameters (the number of classes  $k$ , the classification algorithm and the attributes classified on) are set via the collapsible configuration panel (Fig. 1 (f)). The attributes used in classification will be shown on the attribute uncertainty specification view (Fig. 1 (a)).

The attribute uncertainty specification view uses a parallel coordinates visualization to display user selected attributes. The parallel coordinates plot is known to be suitable for showing multiple data dimensions at once and enables the user to filter each dimension while supporting the identification of correlations. This method is widely used in geo visual analytics systems [13], [34], [35], [64]. Each attribute is one axis, and each spatial unit is represented by a line connecting the values on these axes. Values of an attribute are transformed to be normally distributed, and the bottom matches to the minimum value while the top is the maximum value. A box plot is displayed on each axis to illustrate the value distribution. Lines on the parallel coordinates are colored by the unit’s class label (The legend is shown in Fig. 1 (c)).

On top of each axis, every attribute is associated with an uncertainty configuration area, and by checking the selection box, uncertain attributes are selected for exploring. An attribute uncertainty range and step value can be specified with the plus/minus settings. The user can also specify the uncertainty range per unit by loading a file that describes the uncertainty. The file should include records which contain the unit ID, unit name, uncertainty range and step value. Given the uncertainty of an attribute (or multiple attributes), we run simulations to generate classification results for all possible values in the range at each step interval as described in Section 3.1. For the attributes that have the uncertainty specified, their order can be changed through drag and drop, and this order defines their visual order in the multi-attribute impact profile (Sec. 4.2.2).

### 4.2 Region-based Impact Exploration

In our system, the attribute uncertainty is explored on a per region (spatial unit) basis. Users can explore uncertainty across either one or many attributes to assess the impact on the classification. Multi-attribute analysis is supported, and it is useful for exploring when changes on a combination of multiple attributes may mitigate or exaggerate the effects of uncertainty.

#### 4.2.1 Single Attribute Impact Profile

To analyze the impact of attribute uncertainty, we designed a novel view, the ‘impact river’ (Fig. 2), to visualize the amount of the impact organized by spatial units. In the ‘impact river’ chart, a horizontal axis (Fig. 2 (a)) shows the uncertainty range of the attribute. Each row represents one spatial unit. Each cell of the row encodes the impact of changing that spatial unit’s attribute value with respect to both the variation of information (the upper orange rectangles) and the spatial autocorrelation (the lower blue rectangles). The height and the opacity of the rectangles denote the quantification impact, where Eq. 3 is used for the variation of information and Eq. 7 is used for the spatial autocorrelation. For example, when exploring the uncertainty across an attribute, the cell

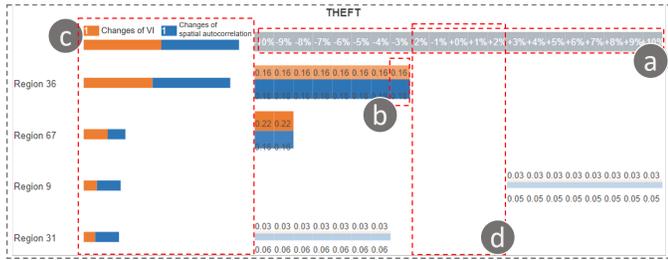


Fig. 2. Illustration of the impact river chart. (a) The horizontal axis shows the uncertainty range of the attribute. Here, the uncertainty range is in the form of a percent change for the base value, and the extremes of the range represent the minimum and maximum value that could be assigned to a particular geographic unit. (b) The cell shows that when decreasing the value of the attribute ‘Theft’ by 3% in *Region*<sub>36</sub>, the impact on the variation of information and spatial autocorrelation are 0.16 and 0.18 respectively. (c) The bars denote the weighted impacts of each spatial unit. (d) The uncertainty range from  $-2\%$  and  $+2\%$  which has no impact on either the variation of information or the spatial autocorrelation.

(Fig. 2 (b)) shows that when decreasing the value of the attribute ‘Theft’ by 3% in *Region*<sub>36</sub>, the quantification of the impacts on the variation of information and spatial autocorrelation are 0.16 and 0.18 (the impact on global spatial autocorrelation) respectively.

To help the analyst identify the spatial units that are most sensitive to uncertainty, the impact river chart sorts all spatial units according to a weighted sum of the impact on the variation of information and the spatial autocorrelation. The analyst can modify the weights by sliding the weight bar on the top of Fig. 2 (c). The length of the two bars for each spatial unit denotes the weighted impacts of each unit. From the impact river chart, the analyst can observe the trend of the classification impact over the range of attribute uncertainty and understand the tolerance of uncertainty over each spatial unit. For example, in Fig. 2 (d), when the value changes within the range of  $-2\%$  and  $+2\%$ , there is no impact on either the variation of information or the spatial autocorrelation. This indicates that the attribute A is not sensitive to small changes, and the resulting map design on the single variable classification, and measures of local spatial autocorrelation, should be robust.

#### 4.2.2 Multi-Attribute Impact Profile

While the impact river chart works well in the single attribute case, in cases such as geodemographic profiling, classification is often performed over multiple attributes. To support the exploration of uncertainty over multiple attributes on the same spatial unit, we employ the use of a multi-layer impact matrix (Fig. 3). Simulations are run for all possible combinations of attribute values in a specified range, and we organize the impact on variation of information and spatial autocorrelation by spatial units and attributes using a matrix view and force-directed layout. For each spatial unit, the order of the layer depends on the attributes order as specified in the uncertainty specification view. At the beginning of the analysis, the impact matrix only shows the simulation results over the first two attributes. Each cell of the matrix matches to a pair of values that the attributes have changed by. Given these changes, the color shows the average amount of the impact over all simulations as the attributes change over their uncertainty range. The inner square (in blue) shows the quantified impact on the spatial autocorrelation, while the outer square (in orange) shows the impact on the variation of information. For example, the cell *a* in Fig. 3 illustrates the impact when the attribute ‘Disturbances’ is

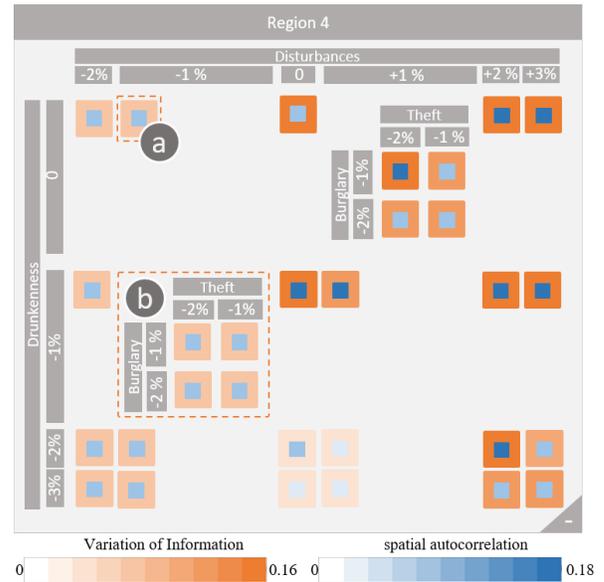


Fig. 3. Visual design for the impact matrix: The inner square (in blue) shows the quantified impact on the spatial autocorrelation, while the outer square (in orange) shows the impact on the variation of information. The darker color refers to larger values. (a) A cell represents the average impact when attribute ‘Disturbances’ is changed by  $-1\%$  and ‘Drunkness’ remains unchanged. (b) A matrix layer shows if attribute ‘Disturbances’ and ‘Drunkness’ both change by  $-1\%$ , what is the impact when attribute ‘Burglary’ and attribute ‘Theft’ changes by  $-2\%$  or  $-1\%$ .

changed by  $-1\%$  and attribute ‘Drunkness’ remains unchanged. To inspect more attributes, the user can click on a cell to expand the impact matrix. The expanded matrix will fix the change of the two outer attributes and visualize the impact of the next two attributes (or one attribute when there is only one left). For example, in the expanded matrix marked as *b* on Fig. 3, ‘Disturbances’ and ‘Drunkness’ are the two outer attributes and fixed as changing by  $-1\%$ , the cells visualize the impact when attribute ‘Burglary’ and ‘Theft’ change by  $-2\%$  or  $-1\%$ . When the number of spatial units under analysis is too large, the analyst can filter the profiles by setting a limited number of impacts on variation of information and spatial autocorrelation. Spatial units that had impacts less than the specified threshold will be filtered out.

#### 4.2.3 Profile Projections

The multi-attribute impact profiles for all spatial units that have a non-zero impact on variation of information or spatial autocorrelation will be projected in the analysis view of the system. We have implemented two types of profile projections to support the mapping between the impact profile to a unit’s geolocation, and the analysis of similar units with respect to their impact quantities. Both projections are implemented by adding position constraints to a force-directed layout method. The implementation has two steps, initialization and adjustment.

**Initialization:** The position of the impact profiles is initialized using the spatial units’ geolocation or 2D coordinates given by PCA (principal component analysis) from a high-dimensional representation of the impact values.

- **The geo-based layout** uses the coordinates corresponding to the unit’s geospatial centroid (the latitude and longitude of the center of the spatial unit).

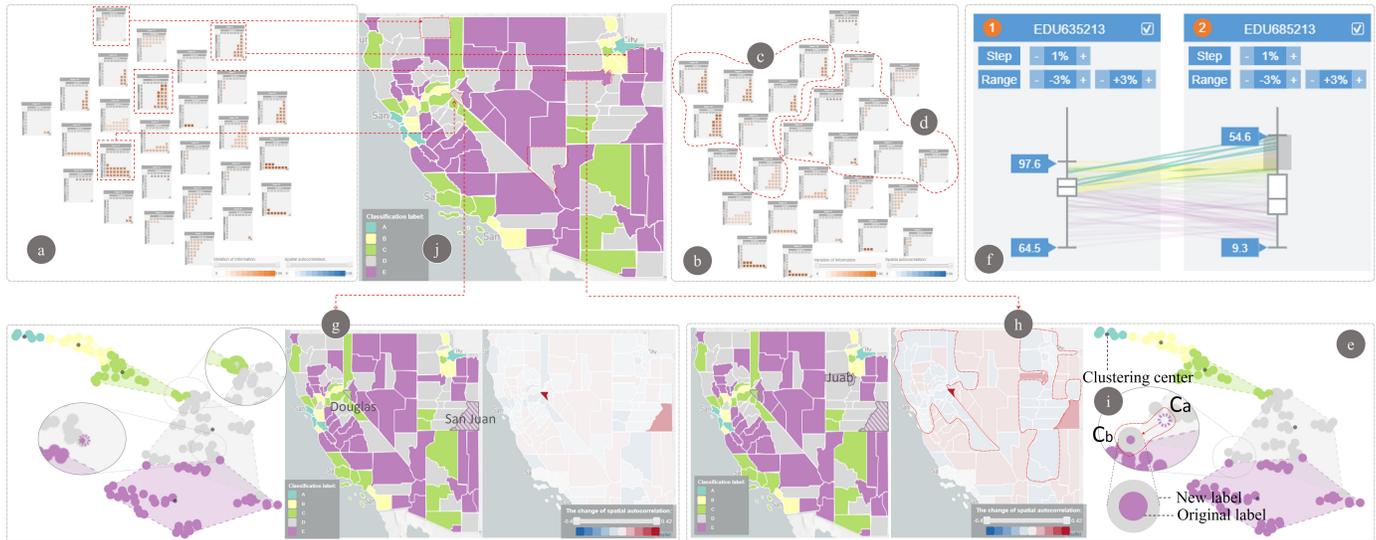


Fig. 4. Illustration of the views in the system. (a) and (b) show the impact matrix in the map-based layout and the PCA-based layout respectively. (c) marks a group of spatial units which all have impacts shown on their bottom-right rectangles, and (d) marks a group of spatial units which have little impact on variation of information and spatial autocorrelation. The t-SNE projection (e) shows the glyph for encoding the comparison of the classification result in a 2-D space. (f) shows the parallel coordinates of attribute ‘EDU635213’ and ‘EDU685213’ and (g) shows the impact that uncertainty in ‘Douglas’ will have, and (h) shows the impact that uncertainty in ‘Jaub’ will have. (i) shows the spatial unit change from  $C_a$  to  $C_b$ .

- **The similarity-based layout** uses the first two principal coordinates given by PCA. We extract the quantified impact of all cells in each impact matrix to obtain a set of high-dimensional vectors. Then, we run the PCA algorithm to convert vectors into PCA space and reduce them to 2D coordinates.

**Adjustment:** After initializing the matrices in 2D space, according to their coordinates from geolocation or PCA projection, these impact profile matrices could suffer from being too sparse or too cluttered. To improve the readability, we employ a force-directed layout approach to iteratively adjust their positions. Specifically, we employ the Dynamic Natural Length Spring algorithm (DNLS) [47] to solve the problem of dynamic drawing overlap.

**Visual Interpretation:** One example of the impact profile projection result is shown in Fig. 4. The relative position of matrices in the Geo-based layout (Fig. 4 (a)) corresponds to their geographical coordinates on the map as illustrated by the red dashed lines. In the PCA-based layout, the matrices that have a similar impact on the variation of information and spatial autocorrelation are placed close together. For example, Fig. 4(b) shows that spatial units in the group  $c$  all have impacts shown on their right rectangles in the matrix. This pattern indicates that these units will have an impact on the classification result when the first attribute being analyzed tends to increase. In contrast to group  $c$ , units in the group  $d$  show a minimal impact on variation of information and spatial autocorrelation when their attribute values change. Note that during the exploration of the matrices, expanding the impact matrix will also lead to overlap. To solve this problem, we perform additional adjustment steps until the position converges.

### 4.3 Simulation-based Impact Exploration

For each attribute value change, our system runs one classification simulation and compares the simulated results to the original classification. In order to support such a comparison, we have

implemented a dual-choropleth map view, a t-SNE projection of classes, and an impact quantity scatter plot.

**Dual-Choropleth Map:** For the exploration of one simulation, our system uses a dual-choropleth map to show the variation of information and the spatial autocorrelation impact for each spatial unit (Fig. 1 (c)). The dual-choropleth contains a classification impact map (left) and a spatial autocorrelation impact map (right). When one simulation is under analysis, the classification impact map colors each spatial unit based on their class label, while the spatial autocorrelation impact map colors each spatial unit based on the magnitude and direction of the change of its local spatial autocorrelation (local Moran’s I).

Specifically, in the scenario that an attribute of spatial unit  $i$  is changed due to some uncertainty, and this change results in that spatial unit  $j$  changing classification labels, the classification map will then highlight unit  $i$  with a wide border and render unit  $j$  with a striped texture where the color of the narrow strips refers to the original classification label while the color of the wide strips refers to the new classification label. Similarly, the spatial autocorrelation impact map will visualize spatial units whose local Moran’s I changes more than a user-defined threshold. Blue indicates a decrease of a unit’s local Moran’s I and red indicates an increase. While the hue of the color is proportional to the magnitude of spatial autocorrelation change, a slider is provided to filter out units with small change.

**t-SNE Classification Projection:** To illustrate how these spatial units change in their attribute space and how some units change classification labels, we use t-SNE projection to reduce selected attributes used in the classification to two dimensions and plot all units on a scatterplot. Each spatial unit is represented as one dot whose color matches the class label. Each class is further shaded by the color to indicate the boundaries of each group.

Given a simulation, this visualization displays the spatial unit with a changed attribute as an enlarged dot and encodes the

impacted units as a special glyph. Fig. 4 (e) shows the visual encoding of the glyph, where the color of the inner circle denotes its original class, and the outer circle represents the new class. Other than spatial units, class centroids are also visualized in this plot. Each centroid is displayed as a black dot in the barycenter of each class. We use animation to visualize the change in classification. This allows the analyst to see how classes change and how one unit moves towards or away from its original class. For example, if an element changes from  $C_a$  to  $C_b$ , we can observe that it moves from the position close to  $C_a$  to somewhere close to  $C_b$  (Fig. 4 (i)).

**Impact Quantity Scatter Plot:** As shown in Fig. 1 (d), to visualize the impact on variation of information and spatial autocorrelation, we plot each simulation on a scatter plot where the x-axis is the variation of information (VI) and the y-axis is the global spatial autocorrelation ( $Impact_s$ ) derived from the local Moran's I. This scatter plot provides an overview of the impact of uncertainty on empirical results.

**Coordinated Interactions:** All visualizations are linked. On the impact profile visualization, if one spatial unit is selected, the unit will be highlighted in the dual-choropleth map by a thick-red boundary. For all values in the uncertainty range, if the corresponding simulation impacts the classification result, a dot will appear on the impact scatter plot. Thus, all simulations for the selected unit will be plotted on the scatter plot to visualize the overall impact of uncertainty. In addition, to visualize the robustness of the spatial unit to attribute uncertainty, the uncertainty range of this unit where no impact is observed will be illustrated with a gray rectangle on the attribute axis. When there is no selected unit, the impact on VI and spatial autocorrelation of all simulations for all spatial units are visualized on the scatter plot.

Linked interactions are enabled when a single simulation (the attribute value is changed by a specific value on a specific spatial unit) is selected. Users can mouse over one simulation on the impact profile visualization or on the impact scatter plot. The changes in attributes will be visualized on the attribute specification parallel coordinates. First, the line associated with the spatial unit will be highlighted at its original location. Then the values of the changed attribute will be used to draw another line indicating the position of the unit after the change. The original values and the new values are all marked on the axis to inform the user of the exact value change. If after this change the unit class flips, the new line will be colored in the new class color. For analyzing the simulation, the corresponding impact of the variation of information and the spatial autocorrelation will be displayed in the dual-choropleth map. The t-SNE view will show the changes in the projection of the data and enlarge the corresponding points of the spatial units. Users can click on the impact profile view to select a particular simulation and this will fix the visualization of the dual-choropleth map and t-SNE view. Also, in the dual-choropleth map, when the user hovers over a spatial unit, the detailed data of the unit will pop up, and its corresponding profile will be highlighted.

## 5 CASE STUDIES

In this section, we will demonstrate how our proposed metrics and visual analytics environment enable the exploration of attribute uncertainty and its impact on choropleth map design and analysis through three case studies. We will start with a one-dimensional

classification case followed by a bivariate classification analysis. We conclude with an example of multivariate classification.

### 5.1 Single Attribute Classification

In our first case study, we use Chicago crime data to illustrate how value changes for one attribute could impact classification results and spatial autocorrelation. This dataset (which can be downloaded at <https://data.cityofchicago.org/>) contains administrative records of crime events in 76 community areas in 2015. There are 148,004 crime events with 10 crime types, among which 38.7% are marked as 'Theft'. Our goal is to develop a choropleth map to illustrate the distribution of 'Theft' events and explore how attribute uncertainty might reshape the choropleth map.

Using the k-means algorithm, we classify the 76 spatial units into 6 classes. We assign each class a color where darker colors indicate more events as shown in Fig. 5 (d). By observing the distribution of these 'Theft' events in the attribute uncertainty specification view, we find that the maximum number of 'Theft' records in one region is 4,182, and 75% of the regions only had 41 to 856 'Theft' events (with a median of 543). To explore the tolerance that this classification would have on uncertainty, we specify the uncertainty range to be -10% to +10% with a step of 1%. Here we note that if uncertainty information is provided, the system would load this for each spatial unit. In this example, uncertainty could arise from miss-classification of events, geocoding errors, etc.

**No Impact Range:** We first look at the sensitivity of all regions, that is, we explore to what degree the uncertainty of this attribute would (or would not) impact the classification results. From the impact river chart (Fig. 5 (a)), we find that 61 regions have no impact on the classification results when their values change between -10% and 10%. Furthermore, we can observe that when 'Theft' events are analyzed, the overall classification result and choropleth map on 6 classes will not be impacted when the record value for one region is off between -2% and +2%.

**Group Patterns on the Classification Impact:** Next, by screening the impact river, we find that the spatial units that may have an impact on the classification results are separated into two groups. In one group, a region tends to impact the classification result when it has more 'Theft' events, while in the other group a region will impact the classification result when it has fewer events. Examining the 'Theft' event counts in these regions on the parallel coordinates plot and the class projections on the t-SNE view, we observe that these regions are at the boundary of two classes. This indicates that they are more prone to changes in classification. Thus, a small amount of turbulence in these near-boundary regions might flip them to another group. We also note that all the regions with a light color, which denotes a low crime rate, will have no impact on both the visual appearance and classification result when their 'Theft' attribute value is changed in the range from -10% to +10%. In other words, these regions have a relatively high tolerance to data uncertainty.

**Most Influential Regions:** After analyzing group patterns with respect to sensitivity, we focus on a few regions which have been identified to be the most influential ones, i.e., units that will impact the classification result with a small change to their attribute values. When the number of 'Theft' event in *Region<sub>57</sub>* drops by 3% (Fig. 5 (c)), this region will be classified into a lighter color group (the

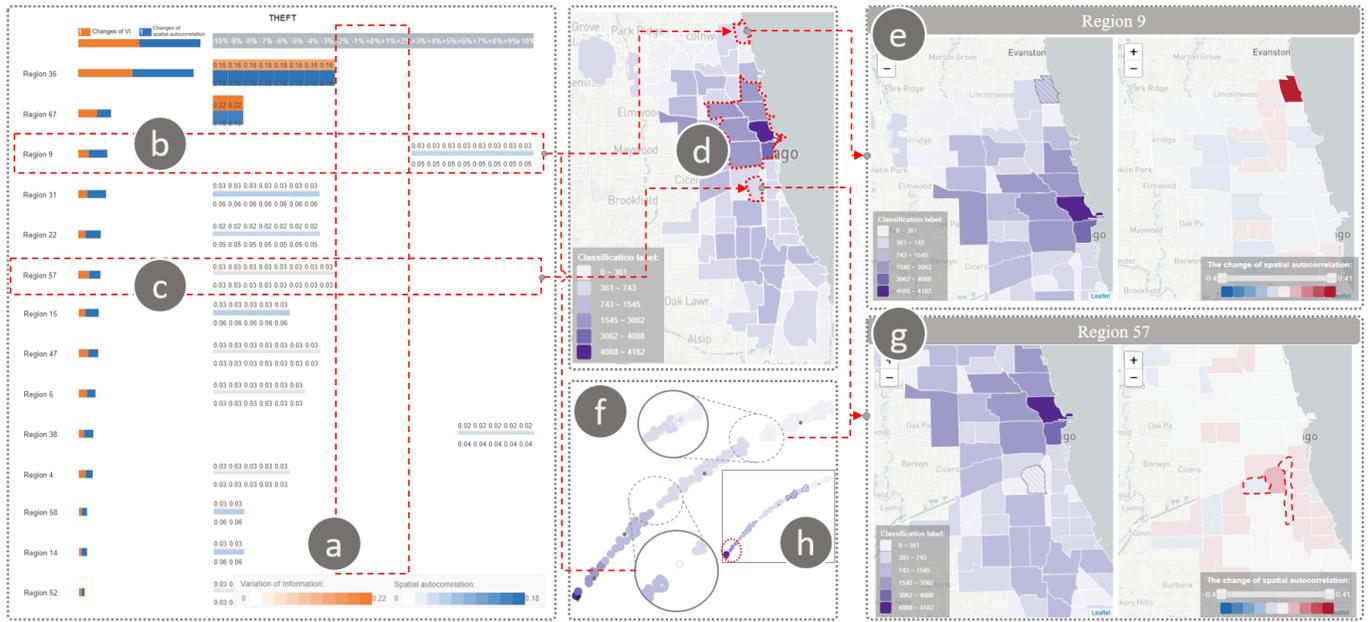


Fig. 5. (a) The impact river shows that when the attribute is changed by  $-2\%$  to  $+2\%$  there is no impact on the classification results. (b) *Region<sub>9</sub>* starts to change class with its ‘Theft’ record increased by  $3\%$ . (c) *Region<sub>57</sub>* will change its class when its ‘Theft’ record decreased by  $3\%$ . (d) The choropleth map illustrates the distribution of ‘Theft’ events. (e) The impact that uncertainty in *Region<sub>9</sub>* will have. When the ‘Theft’ record of *Region<sub>9</sub>* increases by  $3\%$  or more, it will be classified to a darker color class, and therefore increase the local Moran’s I of *Region<sub>9</sub>* and its surrounding areas. (f) The t-SNE view illustrates how *Region<sub>9</sub>* and *Region<sub>57</sub>* change in attribute space. (g) The impact that uncertainty in *Region<sub>57</sub>* will have. When the ‘Theft’ record of *Region<sub>57</sub>* decreases by  $3\%$  or more, it will connect the eastern and western spatial units which creates a larger contiguous region so that will decrease the local Moran’s I of eastern and western regions. (h) The impact of uncertainty in *Region<sub>36</sub>*.

shaded region in Fig. 5 (g)). By analyzing the impact on the classification results, we find that the local Moran’s I of its eastern and western contiguous regions is decreased while its northern neighbor’s local Moran’s I is increased. The reason is that *Region<sub>57</sub>* is reclassified from the northern region’s class to the eastern and western regions’ class. Therefore, *Region<sub>57</sub>* connects the eastern and western regions and creates a larger contiguous region. Next, we investigate the observations around *Region<sub>57</sub>* and find that its surrounding observations do not impact the classification results when their attribute values change in the specified range ( $-10\%$  to  $+10\%$ ). This seems to indicate that *Region<sub>57</sub>* is a critical spatial unit whose accuracy may determine the visualization results of this local area, thus making the requirement of the data accuracy of *Region<sub>57</sub>* a high priority.

Similarly, we can observe *Region<sub>9</sub>* (the shaded region in Fig. 5 (e)), whose class will change when its ‘Theft’ record increases by  $3\%$  (Fig. 5 (b)). Looking at the dual-map view, we find that when the ‘Theft’ record of *Region<sub>9</sub>* increases by  $3\%$ , it will be classified to a darker color class, and therefore detaches from its neighboring regions. As such, the local Moran’s I of *Region<sub>9</sub>* and its surrounding areas will increase. Alternatively, we find that in *Region<sub>36</sub>* if the ‘Theft’ attribute value decreases by  $3\%$ , the classification results will be globally reshaped as shown in Fig. 5 (h). This is due to the attribute value distribution. In the original 6 classes, the two regions with the first and second highest number of ‘Theft’ events each formed a single class. *Region<sub>36</sub>* has the most ‘Theft’ events, and when decreased, is merged into the nearby region which has the second highest ‘Theft’ record. Such a change causes one class to become empty. Therefore, the system will select several regions to form a new class and such re-classification will change the labels of almost all spatial units. In the original classification result, the darkest color class has one region (*Region<sub>36</sub>*) and the second darkest

color class also has one region. Other than these two regions, the remaining regions form 4 classes. After the decrease in *Region<sub>36</sub>*, the darkest class contains two regions since the highest region and the second highest region are regrouped to form one class.

As mentioned by the dataset provider, the addresses of the crime events are reported at the block level, and the crime type, count, and address are not guaranteed to be accurate due to possible mechanical and human errors. Therefore, it is critical to know about the attribute uncertainty while using classification and choropleth maps to analyze these data.

**Generalization:** In order to demonstrate the generalization of the effectiveness of our visual analytics methodology on other datasets, datasets with different distributions are explored. We generated geospatial data with a Normal distribution and heavy-tailed distribution. Then, we ran Jenks optimization and manually assigned these data to regions to obtain maps with large spatial autocorrelation (e.g., the Global Moran’s I of Fig. 6 (a) and Fig. 6 (c) is 0.69 and 0.68 respectively) as well as small spatial autocorrelation (e.g., the Global Moran’s I of Fig. 6 (b) and Fig. 6 (d) is  $-0.20$  and  $-0.25$  respectively). Next, we specify the uncertainty range to be  $-10\%$  to  $+10\%$  with a step of  $1\%$  and begin our analysis. We observed that the results show the same characteristics as in the real dataset. For example, the red rectangles in Fig. 6 indicate that all regions have a ‘no impact range’, and the regions can be clearly classified into two groups. In one group, a region tends to impact the classification result when it has a larger data value, while in the other group, a region will impact the classification result when it has a smaller data value. Comparing the impact river of the Normally distributed data and the heavy-tailed distributed data, we found that the regions are more sensitive to the data uncertainty when they follow a Normal distribution. The reason is that the gap between the classes is small when the data distribution

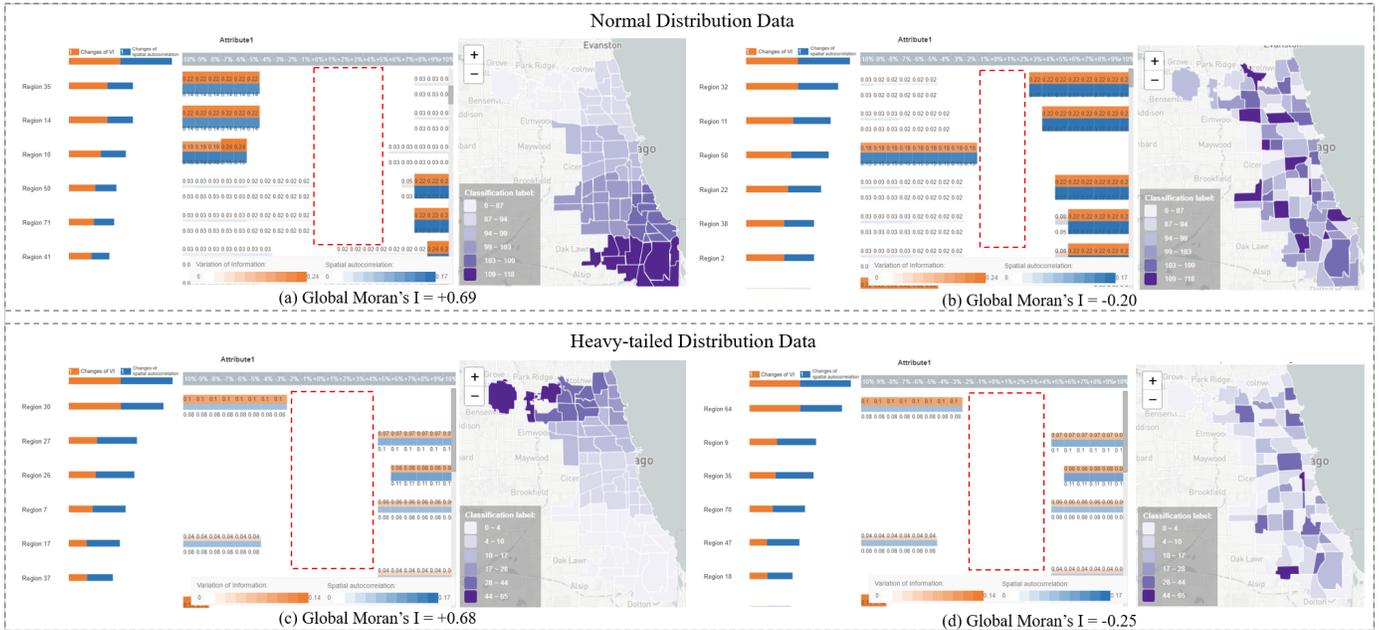


Fig. 6. Single attribute classification results with (a) Normal distribution data and large Global Moran's I (b) Normal distribution data and small Global Moran's I (c) heavy-tailed distribution data and large Global Moran's I, and (d) heavy-tailed distribution data and small Global Moran's I.

is Normal. Different spatial autocorrelation values do not seem to drastically impact the results of the analysis. For example, in Fig. 6 (c) and (d), despite having different Global Moran's I, the gap in the analysis results is negligible. Further studies on data distributions on uncertainty impact are left as future work.

### 5.2 Bivariate Classification

To explore the impact of attribute uncertainty on classification and choropleth maps when performing bivariate classification, we used the U.S. Census data ([https://geodacenter.github.io/data-and-lab/county\\_election\\_2012\\_2016-variables/](https://geodacenter.github.io/data-and-lab/county_election_2012_2016-variables/)) and extracted the 118 counties in the southwestern United States (Covering 4 states: CA, AZ, UT, and NV). Among all available attributes, we selected two educational variables, 'EDU635213' and 'EDU685213'. 'EDU635213' represents the percentage of high school graduates or higher education populations over the age of 25, and 'EDU63521' represents the portion of people who have bachelor's degrees or higher. Our goal was to develop a choropleth map to explore these educational variables.

**Attribute Correlation:** We expect that the two educational attributes should be highly correlated because a region with high education levels should have both higher rates of secondary education and college degrees. However, when we look at the parallel coordinates plot (Fig. 4 f), our expectation is only partially supported. We find that in the regions with the lowest rate of secondary education (colored in pink), the level of higher education is, unsurprisingly, the lowest, and the regions with the highest levels of education also have the highest secondary education rates (e.g., Sacramento and San Francisco). However, there are some regions that have a high rate of secondary education but a low rate of higher education (e.g., Box Elder County in UT). We run k-means and generate a 5-class choropleth map as shown in Fig. 4 (j). We specify the uncertainty range of both secondary and higher

education rates as -3% to +3%, with a step of 1%.

**Group Patterns on the Impact Profiles:** The region-based impact profiles are first visualized using the geo-based layout. In each impact profile matrix, the rows represent the changed percentage of the secondary education rate, and the columns represent the changed percentage of the higher educational rate. We find that there are only a few regions in the Central and South area that are influential in the map design. However, regions in the Northeast and Northwest have high sensitivity to uncertainty, and some of the regions greatly impact the classification results when their attribute values change. Thus, counties in the northeastern area might need more attention to data uncertainty.

Turning to the similarity-based layout of the impact profiles (Fig. 4 (b)), we find that increasing the value of the data will have more impact than decreasing the value since most changes are visualized on the lower-right corner of each impact matrix. We also find that the colors of some matrices change with each column, which means that the regions are more sensitive to the uncertainty of the secondary education attribute than to the higher education attribute. We carefully investigate these regions on the map and find that these regions are all colored pink, which means their secondary education rate is low and determines their classification label. There are other matrices that change colors with each row, this is because their secondary education is already widespread, and the gap in higher education rate is larger than in secondary education. Therefore, the classification results are influenced by the level of higher education.

**Most Sensitive Regions:** Next, we filter the regions by their impact using the sliders and focus on the regions that have a large impact on the classification results. One of these regions is Douglas County (*Region*<sub>39</sub> as shown in Fig. 4 (g)). When its higher educational rate increases by 1%, the classification results will have a significant change throughout the regions. Looking at the t-SNE view and the

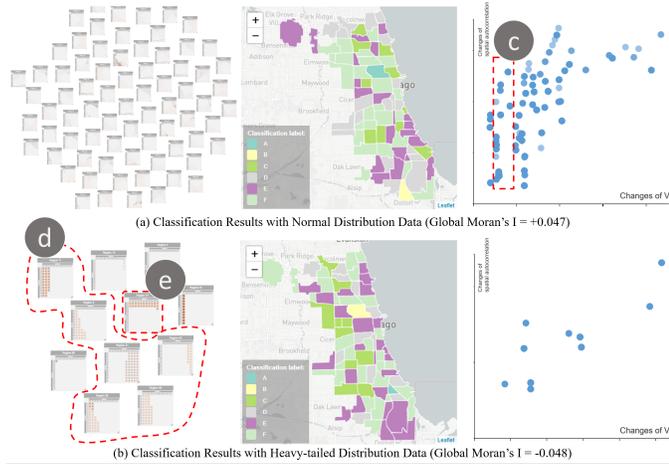


Fig. 7. Illustration of the bivariate classification results of data with (a) Normal distribution and (b) heavy-tailed distribution. (c) The impact quantity scatter plot shows that with the similar variation of information, the impact on spatial autocorrelation can be different. (d) The regions that are sensitive to uncertainty in  $attribute_2$ . (e) The regions that are sensitive to uncertainty in  $attribute_1$ .

dual-map view, we find that when the value of ‘Douglas’ increases, it will be assigned to a group with a higher level of education. Meanwhile, another county, ‘San Juan’, will also be assigned to a higher level education group since their education rates are at the ceiling of their original group. The visual spatial impact covers the entire area under classification. As illustrated in Fig. 4 (g), all areas experience a change in their local Moran’s I. This means Douglas County drives the map classification results.

Similarly, Juab County in Utah ( $Region_{101}$  as shown in Fig. 4 (h)) is sensitive to uncertainty in the secondary education rate. By increasing the level of secondary education rate by 1%, Juab County will be assigned to a class with a higher educational level. After investigating the impact on spatial autocorrelation, we find that the local Moran’s I of all regions have changed. Specifically, the northwest and northeast regions had their local Moran’s I decreased because some regions changed their class and connected with the surrounding regions.

**Generalization:** To explore a generalization of bivariate classification effects, we generated sample two-dimensional datasets with a Normal distribution and heavy-tailed distribution. We assigned these data to spatial areas and run k-means to generate 6-class choropleth maps. The Global Moran’s I of classification results with Normal distribution data (as shown in Fig. 7 (a)) and heavy-tailed distribution data (as shown in Fig. 7 (b)) are +0.047 and -0.048 respectively. We specify the uncertainty range to be -10% to +10% with a step of 2% and begin our analysis. As shown in Fig. 7, in each impact profile matrix, the columns represent the changed percentage of  $attribute_1$ , and the rows represent the changed percentage of  $attribute_2$ . Looking at the impact quantity scatter plot, we found that even with the similar variation of information, the impact on spatial autocorrelation can be different, especially in the Normally distributed data (as shown in Fig. 7 (c)). Next, we observed that the impact profiles show group patterns. For example, the regions in Fig. 7 (d) are sensitive to uncertainty in  $attribute_2$ , while the region corresponding to Fig. 7 (e) is sensitive to uncertainty in  $attribute_1$ . We also found that there are more impact profiles in Fig. 7 (a) than Fig. 7 (b), which indicates that the areas are more sensitive to data uncertainty when their values

are Normally distributed. The reason is that observations in the Normal distribution are closer so the gap in the classification result is smaller than in the heavy-tailed distribution.

### 5.3 Multivariate Classification

For complex classification problems that have multiple attributes, relevant analysis includes the identification of which attributes are the least and most significant to the classification result, which regions are most sensitive to uncertainty with respect to their class labels, and what might be misinterpreted if the data is uncertain.

To demonstrate the effectiveness of our system when analyzing the impact that uncertainty can have when dealing with multiple attributes, we run a classification of 6 classes for the following six attributes: education (‘EDU635213’, ‘EDU685213’), median household income (‘INC110213’), the percentage of foreign-born persons (‘POP645213’), the difference between the number of votes for Republican and Democratic candidates in 2012 (‘diff2012’), and the percentage of persons below poverty level (‘PVY020213’). The classification results are shown in Fig. 1 (c). We find that the most developed areas (e.g., San Francisco and Los Angeles) are colored in blue and purple. These areas have a high percentage of foreign-born persons and a high educational level. In contrast, the orange and pink groups (e.g., the northern area of California and the southern area of Nevada) have lower educational attainment and lower median household incomes.

**Influential Attributes:** We selected four attributes to explore the impact of uncertainty, and set the uncertainty range of ‘POP645213’ (foreign-born population), ‘EDU685213’ (education), ‘INC110213’ (household income), and ‘PVY020213’ (persons below poverty level) from -5% to +5% with a step of 2%. The initial analytic order of attributes is foreign-born population, education, household income, and persons below the poverty level. We filter out regions that have a small impact by setting the threshold of classification, quantified by the variation of information, as ‘0.04’. Looking at the impact profiles of the regions (Fig. 1 (b)), we find that almost all of the matrices (except  $Region_{44}$  and  $Region_{45}$ ) change their colors along the columns, which indicates that the uncertainty of the first-order attribute ‘foreign-born population’ has a broad impact on the classification results.

**Impacts from Other Attributes:** Looking at San Benito County ( $Region_{58}$  as shown on Fig. 1 (g)), we find a large number of cells colored in the impact matrix centered at the bottom left. This means that when the value of ‘foreign-born population’ decreases and the value of ‘education’ increases, San Benito County will impact the classification result. We expand this impact matrix to explore the impact on the next two attributes. As illustrated in Fig. 1 (h), the sub-matrix of San Benito County changes its color according to the value change of the third attribute ‘household income’. To summarize, the uncertainty of the foreign-born population, education and household income of San Benito County impact the classification.

Next, we select one of these simulations to analyze the local impact on these regions. We find that when the attributes of San Benito County changed, some other regions are impacted and switch their class from gray to blue or from blue to other classes. As illustrated in Fig. 1 (i), the local Moran’s I increased in the northeast and decreased in the northwest. Thus, it is important to require accurate attribute values for San Benito County because

the overall classification seems to be very sensitive to its value change. There are also some regions in gray and blue that are easily impacted so analysts may need to pay more attention to them. In addition, the attribute value for household income also has a significant impact on the clustering results.

In order to further explore the relationship between these attributes in terms of their classification impact, we move ‘household income’ to second place (from fourth place) in the list of analytic attributes. We find that the impact of many regions in the matrices is shown as a triangle. In some regions, when the household income and the number of foreign-born increase, it will impact the classification results. However, in Carbon County (*Region*<sub>102</sub> as shown in Fig. 1 (j)), only the uncertainty of household income will impact the classification result. We click the impact matrix of Carbon County and expand the matrix view to explore it in detail. We find that the impact of the area on the clustering results increases when the attribute value of household income and poverty rates change. We also find that Carbon County is changed into the gray cluster, and it will impact many other regions switching their colors to blue. As a result, the local Moran’s *I* of the northeastern areas have increased while the western areas decreased.

Finally, we select ‘persons below poverty level’ as the second order analytic attribute. Looking at the impact profiles, we find that except for *Region*<sub>102</sub>, *Region*<sub>7</sub>, *Region*<sub>44</sub> and *Region*<sub>34</sub>, the colors of the matrices do not change along the rows. This indicates that the classification results are not sensitive to the uncertainty range of the persons below poverty level attribute.

## 6 CONCLUSION AND FUTURE WORK

In this work, we proposed a visual analytics methodology to help map designers and analysts explore the impact that uncertainty of data attributes will have on both map classification results and spatial autocorrelation. Our system consists of a parallel coordinates-based attribute specification view, an impact river visualization for single attribute classification analysis, an impact profile visualization for multiple attributes classification analysis, a dual-choropleth map and t-SNE-based visualization to illustrate the class change of spatial units and the visual impact. We demonstrated the possible findings using this system to explore uncertain data through three concrete examples.

From our case studies, we demonstrate how our system can be used to show the impact of uncertainty on map classification and spatial analysis. We also find that, in many cases, it was the boundary units (between two classes) that tend to flip their classes. However, this may be caused by a value change in another unit, not necessarily the class-changing unit itself. Furthermore, some class flips of multiple spatial units are co-instantaneous. Therefore the analyst should consider the integral effect. We also observed that the influence on the map classification result is different among attributes. By looking at the pattern on the impact profiles and changing the attribute order, the analyst can learn how the uncertainty of some attributes might cause more significant changes on map classification than other attributes. Finally, when multiple attributes are analyzed together, the impact of different attributes might counteract each other. In this cases, these attributes could both be sensitive to classification.

However, our system also has several limitations. First, when the data dimension and the number of spatial units increase, the number of simulations will become large, and our system will suffer from an increased computational cost. A possible solution could

be to preprocess the data with a preset uncertainty range and step. For the impact matrix view, as the number of selected attributes and uncertainty range increases, it is inconvenient to show the entire matrix by clicking the rectangles one by one. Furthermore, the views designed may not be optimal for identifying changes, and future work could explore novel design alternatives. Second, when the analyst uses this tool to explore data, some findings and discoveries might also be related to the clustering algorithm used in that particular scenario. Finally, our system analyzes the impact due to a single spatial unit; however, future work should consider cascading effects when multiple units change simultaneously.

## ACKNOWLEDGMENTS

Wei Chen and Zhaosong Huang are supported by National 973 Program of China (2015CB352503) and National Natural Science Foundation of China (61772456, 61761136020). Zhaosong Huang holds a China Scholarship Council Studentship at Arizona State University. Ross Maciejewski and Yafeng Lu are supported by the National Science Foundation (1350573) and by the U.S. Department of Homeland Security under Grant Award 2017-ST-061-QA0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

## REFERENCES

- [1] J. C. Aerts, K. C. Clarke, and A. D. Keuper. Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science*, 30(3):249–261, 2003.
- [2] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on Visual Analytics Science and Technology*, pp. 3–10. IEEE, 2009.
- [3] L. Anselin. *Spatial econometrics: methods and models*, vol. 4. Springer Science & Business Media, 1988.
- [4] L. Anselin. Local indicators of spatial association-LISA. *Geographical Analysis*, 27(2):93–115, 1995.
- [5] M. P. Armstrong, N. Xiao, and D. A. Bennett. Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers*, 93(3):595–623, 2003.
- [6] D. I. Ashby and P. A. Longley. Geocomputation, geodemographics and resource allocation for local policing. *Transactions in GIS*, 9(1):53–72, 2005.
- [7] P. Berkhin. A survey of clustering data mining techniques. In *Grouping Multidimensional Data*, pp. 25–71. Springer, 2006.
- [8] D. Birant and A. Kut. ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.
- [9] A. M. Bisantz, D. Cao, M. Jenkins, P. R. Pennathur, M. Farry, E. Roth, S. S. Potter, and J. Pfautz. Comparing uncertainty visualizations for a dynamic decision-making task. *Journal of Cognitive Engineering and Decision Making*, 5(3):277–293, 2011.
- [10] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*, pp. 3–27. Springer, 2014.
- [11] C. A. Brewer and L. Pickle. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4):662–681, 2002.
- [12] A. Brolese and S. Huf. Visualizing uncertainty to improve operators’ spatial proximity judgments in uncertain surroundings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, pp. 294–298. Sage Publications Sage CA: Los Angeles, CA, 2006.
- [13] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Visual Analytics Science and Technology*, pp. 27–34. IEEE, 2010.
- [14] D. C. Cliburn, J. J. Feddema, J. R. Miller, and T. A. Slocum. Design and evaluation of a decision support system in a water balance application. *Computers & Graphics*, 26(6):931–949, 2002.

- [15] M. Correll, D. Moritz, and J. Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 642. ACM, 2018.
- [16] J. Cox and M. Lindell. Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, 3(2), 2013.
- [17] E. Cromley and R. Cromley. An analysis of alternative classification schemes for medical atlas mapping. *European Journal of Cancer*, 32(9):1551–1559, 1996.
- [18] S. Deitrick. Evaluating implicit visualization of uncertainty for public policy decision support. In *Proceedings AutoCarto 2012*, p. available online.
- [19] S. Deitrick and R. Edsall. The influence of uncertainty visualization on decision making: An empirical evaluation. In *Progress in Spatial Data Handling*, pp. 719–738. Springer, 2006.
- [20] I. Drecki. Visualisation of uncertainty in geographical data. *Spatial Data Quality*, pp. 140–159, 2002.
- [21] P. F. Fisher. Visualizing uncertainty in soil maps by animation. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 30(2-3):20–27, 1993.
- [22] A. S. Fotheringham and D. W. Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.
- [23] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.
- [24] C. E. Gehlke and K. Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A):169–170, 1934.
- [25] N. D. Gershon. Visualization of fuzzy data using generalized animation. In *IEEE Conference on Visualization*, pp. 268–273. IEEE, 1992.
- [26] A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206, 1992.
- [27] T. Gschwandtner, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):539–548, 2016.
- [28] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. *Information systems*, 25(5):345–366, 2000.
- [29] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.
- [30] G. B. Heuvelink, J. D. Brown, and E. E. van Loon. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, 21(5):497–513, 2007.
- [31] Y. Hu, S. G. Kobourov, and S. Veeramoni. Embedding, clustering and coloring for dynamic maps. In *Pacific Visualization Symposium*, pp. 33–40. IEEE, 2012.
- [32] J. Jeffers. A basic subroutine for Geary’s contiguity ratio. *Journal of the Royal Statistical Society.*, 22(4):299–302, 1973.
- [33] G. F. Jenks. The data model concept in statistical mapping. *International Yearbook of Cartography*, 7:186–190, 1967.
- [34] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. ipca: An interactive system for pca-based visual analytics. In *Computer Graphics Forum*, vol. 28, pp. 767–774. Wiley Online Library, 2009.
- [35] J. Johansson, P. Lung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *Information Visualization*, pp. 125–132. IEEE, 2005.
- [36] J. Kardos, G. L. Benwell, and A. B. Moore. Assessing different approaches to visualise spatial and attribute uncertainty in socioeconomic data using the hexagonal or rhombus (hor) trustee. *Computers, Environment and Urban Systems*, 31(1):91–106, 2007.
- [37] B. Kelly. Review of unclassed choropleth mapping. *Cartographic Perspectives*, (86):30–35, 2017.
- [38] C. Kinkeldey, A. M. MacEachren, M. Riveiro, and J. Schiewe. Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science*, 44(1):1–21, 2017.
- [39] C. Kinkeldey, A. M. MacEachren, and J. Schiewe. How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal*, 51(4):372–386, 2014.
- [40] C. Kinkeldey, J. Mason, A. Klippel, and J. Schiewe. Evaluation of Noise Annotation Lines: Using Noise to Represent Thematic Uncertainty in Maps. *Cartography and Geographic Information Science*, 41(5):430–439, 2014.
- [41] H. Koo, Y. Chun, and D. A. Griffith. Geovisualization of attribute uncertainty. In *Proceedings of the 13th Int. Conference on GeoComputation, The University of Texas at Dallas, Richardson, Texas, USA*, pp. 230–236, 2015.
- [42] H. Koo, Y. Chun, and D. A. Griffith. Optimal map classification incorporating uncertainty information. *Annals of the American Association of Geographers*, 107(3):575–590, 2017.
- [43] E. Koua. Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets. *Proceedings of 21st International Cartographic Renaissance*, pp. 1694–1702, 2003.
- [44] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):142–151, 2018.
- [45] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16–22. ACM, 1999.
- [46] M. Leitner and B. P. Buttenfield. Guidelines for the display of attribute certainty. *Cartography and Geographic Information Science*, 27(1):3–14, 2000.
- [47] W. Li, P. Eades, and N. Nikolov. Using spring algorithms to remove node overlapping. In *Proceedings of the Asia Pacific Symposium on Information Visualisation*, vol. 45, pp. 131–140. Australian Computer Society, Inc., 2005.
- [48] Z. Lin, M. R. Lyu, and I. King. Matchsim: A novel neighborhood-based similarity measure with maximum neighborhood matching. In *Proceedings of the 18th ACM conference on Information and Knowledge Management*, pp. 1613–1616. ACM, 2009.
- [49] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind. *Geographic information systems and science*. John Wiley & Sons, 2005.
- [50] H. MacDonald. The American Community Survey: Warmer (more current), but fuzzier (less precise) than the decennial census. *Journal of the American Planning Association*, 72(4):491–503, 2006.
- [51] A. M. MacEachren. Visualizing uncertain information. *Cartographic Perspectives*, (13):10–19, 1992.
- [52] A. M. MacEachren, C. A. Brewer, and L. W. Pickle. Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A*, 30(9):1547–1561, 1998.
- [53] D. Manley. Scale, aggregation, and the modifiable areal unit problem. In *Handbook of Regional Science*, pp. 1157–1171. Springer, 2014.
- [54] M. Meilă. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines*, pp. 173–187. Springer, 2003.
- [55] M. Meilă. Comparing clusterings an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.
- [56] M. Meilă and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1-2):9–29, 2001.
- [57] M. Monmonier. *How to lie with maps*. University of Chicago Press, 2018.
- [58] L. Nadav-Greenberg, S. L. Joslyn, and M. U. Taing. The effect of uncertainty visualizations on decision making in weather forecasting. *Journal of Cognitive Engineering and Decision Making*, 2(1):24–47, 2008.
- [59] J. K. Nelson and C. A. Brewer. Evaluating data stability in aggregation structures across spatial scales: Revisiting the modifiable areal unit problem. *Cartography and Geographic Information Science*, 44(1):35–50, 2017.
- [60] T. S. Newman and W. Lee. On visualizing uncertainty in volumetric data: techniques and their evaluation. *Journal of Visual Languages & Computing*, 15(6):463–491, 2004.
- [61] L. W. Pickle. *Atlas of united states mortality*. National Center for Health Statistics, Centers for Disease Control and Prevention, US Dept. of Health and Human Services, 1996.
- [62] M. Polczynski and M. Polczynski. Using the k-means clustering algorithm to classify features for choropleth maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 49(1):69–75, 2014.
- [63] K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. Visualizing summary statistics and uncertainty. *Proceedings of Eurovis, Computer Graphics Forum*, 29(3):823–831, 2010.
- [64] H. Qu, W.-Y. Chan, A. Xu, K.-L. Chung, K.-H. Lau, and P. Guo. Visual analysis of the air pollution problem in Hong Kong. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1408–1415, 2007.
- [65] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
- [66] D. Retchless. Mapping climate change uncertainty: Effects on risk perceptions and decision making. In *AGU Fall Meeting Abstracts*, 2012.
- [67] M. Riveiro, T. Helldin, and G. Falkman. Influence of meta-information on decision-making: Lessons learned from four case studies. In *IEEE*

*International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, pp. 14–20. IEEE, 2014.

- [68] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 2009.
- [69] W. Shi. *Principles of modeling uncertainties in spatial data and spatial analyses*. CRC press, 2009.
- [70] A. D. Singleton and P. A. Longley. Geodemographics, visualisation, and social networks in applied geography. *Applied Geography*, 29(3):289–298, 2009.
- [71] A. Slingsby, J. Dykes, and J. Wood. Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2545–2554, 2011.
- [72] S. E. Spielman, D. Folch, and N. Nagle. Patterns and causes of uncertainty in the american community survey. *Applied Geography*, 46:147–157, 2014.
- [73] M. Sun, D. Wong, and B. Kronenfeld. A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. *Cartography and Geographic Information Science*, 44(3):246–258, 2017.
- [74] M. Sun and D. W. Wong. Incorporating data quality information in mapping american community survey data. *Cartography and Geographic Information Science*, 37(4):285–299, 2010.
- [75] M. Sun, D. W. Wong, and B. J. Kronenfeld. A classification method for choropleth maps incorporating data reliability information. *The Professional Geographer*, 67(1):72–83, 2015.
- [76] N. Tagashira and A. Okabe. The modifiable areal unit problem, in a regression model whose independent variable is a distance from a predetermined point. *Geographical Analysis*, 34(1):1–20, 2002.
- [77] W. R. Tobler. Choropleth maps without class intervals? *Geographical analysis*, 5(3):262–265, 1973.
- [78] C. Traun and M. Loidl. Autocorrelation-based reclassification—a self-calibrating classification approach for choropleth maps explicitly considering spatial autocorrelation. *International Journal of Geographical Information Science*, 26(5):923–939, 2012.
- [79] M. Tucci and A. Giordano. Positional accuracy, positional uncertainty, and feature change detection in historical maps: Results of an experiment. *Computers, Environment and Urban Systems*, 35(6):452–463, 2011.
- [80] T. Viard, G. Caumon, and B. Levy. Adjacent versus coincident representations of geospatial uncertainty: Which promote better decisions? *Computers & Geosciences*, 37(4):511–520, 2011.
- [81] T. Von Landesberger, F. Brodtkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. Mobilitygraphs: Visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):11–20, 2016.
- [82] R. Wei and T. H. Grubestic. An alternative classification scheme for uncertain attribute mapping. *The Professional Geographer*, 69(4):604–615, 2017.
- [83] N. Xiao, C. A. Calder, and M. P. Armstrong. Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science*, 21(2):121–144, 2007.
- [84] Y. Zhang, W. Luo, E. A. Mack, and R. Maciejewski. Visualizing the impact of geographical variations on multivariate clustering. In *Computer Graphics Forum*, vol. 35, pp. 101–110. Wiley Online Library, 2016.
- [85] Y. Zhang and R. Maciejewski. Quantifying the visual impact of classification boundaries in choropleth maps. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):371–380, 2017.
- [86] F. Zhou, J. Li, W. Huang, Y. Zhao, X. Yuan, X. Liang, and Y. Shi. Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data. In *Pacific Visualization Symposium*, pp. 128–135. IEEE, 2016.



**Zhaosong Huang** received his B.S. in computer science from the Shandong University, China in 2015. He is currently a PhD student in the College of Computer Science and Technology at the Zhejiang University of State Key Lab of CAD&CG. His research interests include visualization and visual analysis of urban data.



**Yafeng Lu** is a Postdoctoral Research Associate in the School of Computing, Informatics and Decision Systems Engineering at Arizona State University. Her current research interests are in data visualization and predictive analytics.



**Elizabeth Mack** is an Associate Professor in the Department of Geography at Michigan State University where she teaches courses in economic geography. Dr. Macks research program evaluates the impact of information and communications technologies (ICTs) on the development trajectory of regional economies. This research program includes broadband infrastructure deployment policy issues as well as the impacts of broadband on business location. Recently, Dr. Macks work is focused on understanding entrepreneurial ecosystems, water affordability, and spatial data uncertainty. For more information, please refer to <http://www.elizabethmack.com>



**Wei Chen** is a professor at the State Key Lab of CAD&CG, Zhejiang University. His research interests is on visualization and visual analysis, and has published more than 30 IEEE/ACM Transactions and IEEE VIS papers. He actively served as guest or associate editors of the IEEE Transactions on Visualization and Computer Graphics, the IEEE Transactions on Intelligent Transportation Systems, and Journal of Visualization. For more information, please refer to <http://www.cad.zju.edu.cn/home/chenwei/>



**Ross Maciejewski** is an Associate Professor at Arizona State University in the School of Computing, Informatics & Decision Systems Engineering. His primary research interests are in the areas of geographical visualization and visual analytics focusing on public health, dietary analysis, social media, criminal incident reports, and the food-energy-water nexus. For more information, visit <http://vader.lab.asu.edu>.