

VISUALIZING AND MODELING SPATIAL DATA UNCERTAINTY

by

Hyeongmo Koo



APPROVED BY SUPERVISORY COMMITTEE:

Yongwan Chun, Chair

Daniel A. Griffith

Denis J. Dean

May Yuan

Copyright 2018

Hyeongmo Koo

All Rights Reserved

This dissertation is dedicated to my wife, Miyeon Son.

VISUALIZING AND MODELING SPATIAL DATA UNCERTAINTY

by

HYEONGMO KOO, BA, MA

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

GEOSPATIAL INFORMATION SCIENCES

THE UNIVERSITY OF TEXAS AT DALLAS

May 2018

ACKNOWLEDGMENTS

I offer my deep gratitude to my dissertation committee chair, Dr. Yongwan Chun for his consistent support and guidance throughout the process of completing this work. I am forever grateful to Dr. Daniel A. Griffith, who permitted me the opportunity to work with him on my Ph.D. studies. I would like to extend my sincere appreciation to Dr. Dean J. Denis and Dr. May Yuan for their supports and being dissertation committee members. They also inspired me to broaden my work. Lastly, many thanks to all of those, especially my colleagues Monghyeon, Lan, and Nick, who supported me in any respect during the completion of my doctoral studies.

March 2018

VISUALIZING AND MODELING SPATIAL DATA UNCERTAINTY

Hyeongmo Koo, PhD
The University of Texas at Dallas, 2018

Supervising Professor: Yongwan Chun

This dissertation extends the understanding of spatial data uncertainty, which inevitably exists in any process of Geographic Information Sciences involving measuring, representing, and modeling the world. This dissertation consists of three specific sub-topics in visualizing and modeling spatial data uncertainty. First, a framework for attribute uncertainty visualization is suggested based on bivariate mapping techniques, and this framework is implemented in a popular GIS environment. The framework and implementation support many visual variables that have been investigated in the literature. This research outcome can provide flexibility to enhance communication and visualization effectiveness for uncertainty visualization. The second sub-topic is a development of optimal map classification methods by simultaneously considering attribute estimates and their uncertainty. This study expands the discussion of constructing an optimal map classification result in which data uncertainty is incorporated in a map classification process. This method utilizes a shortest path problem in an acyclic network based on dissimilarity measures with various cost and objective functions. Finally, modeling positional uncertainty acquired through street geocoding is investigated to understand potential factors of the uncertainty and then to identify impacts of the uncertainty on spatial analysis results. This study accounts for spatial autocorrelation among

geocoded points in a modeling process, which has been barely included in this type of modeling. This research has contributions to increasing explanation and to extending geocoding uncertainty modeling by suggesting additional covariates and considering spatial autocorrelation.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
ABSTRACT.....	vi
LIST OF FIGURES	x
LIST OF TABLES	xi
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 GEOVISUALIZING ATTRIBUTE UNCERTAINTY OF INTERVAL AND RATIO VARIABLES: A FRAMEWORK AND AN IMPLEMENTATION FOR VECTOR DATA	5
Abstract	6
2.1 Introduction.....	7
2.2 Related work	9
2.3 A framework for uncertainty visualization based on bivariate mapping.....	12
2.4 An implementation of uncertainty visualization.....	15
2.5 An application.....	19
2.5.1 Overlaid Symbols on a Choropleth Map (OSCM).....	20
2.5.2 Coloring Properties on a Proportional Symbol map (CPPS)	22
2.5.3 Composite Symbols (CS).....	24
2.6 Summary and conclusions	25
CHAPTER 3 OPTIMAL MAP CLASSIFICATION INCORPORATING UNCERTAINTY INFORMATION	27
Abstract	28
3.1 Introduction.....	29
3.2 Literature Review.....	31
3.3 Method	35
3.3.1 Optimal Classification.....	35
3.3.2 Cost Functions.....	37
3.3.3 Objective functions and formulations	39
3.3.4 An Optimal Map Classification Implementation	42

3.4	A Comparative Application	43
3.5	Conclusions and limitations	53
CHAPTER 4 MODELING POSITIONAL UNCERTAINTY ACQUIRED THROUGH STREET GEOCODING		57
	Abstract	58
4.1	Introduction	59
4.2	Literature review	61
4.3	Methods	64
	4.3.1 Data	64
	4.3.2 A geocoding process and positional uncertainty quantification	66
	4.3.3 Distribution of geocoding positional uncertainty	67
	4.3.4 Regression model specifications	68
4.4	Results	71
	4.4.1 Descriptive statistics	71
	4.4.2 Distribution fitting results	74
	4.4.3 Regression results	78
	4.4.4 Spatial regression model results	80
4.5	Conclusions	83
CHAPTER 5 CONCLUDING REMARKS		86
APPENDIX A THE GRAPHIC USER INTERFACE		90
APPENDIX B BOOTSTRAPPING RESULTS		92
REFERENCES		96
BIOGRAPHICAL SKETCH		107
CURRICULUM VITAE		108

LIST OF FIGURES

Figure 2-1. Menu interface of the attribute uncertainty geovisualization extension	18
Figure 2-2. Study Area.....	20
Figure 2-3. OSCM for 2013 median age and its coefficient of variation by census block group in Plano, TX	22
Figure 2-4. Coloring properties on proportional symbol maps for the 2013 number of housing units and its coefficient of variation by census block group in Plano, TX	23
Figure 2-5. Composite symbol maps for the 2013 number of housing units with 99 percent confidence intervals by census block group in Plano, TX.....	25
Figure 3-1. An example of an acyclic network with five nodes	37
Figure 3-2. The GUI of the implemented tool	43
Figure 3-3. Classification results using the natural breaks and class separability methods	47
Figure 3-4. Optimal classification results by the class separability measure	48
Figure 3-5. Optimal classification result maps using the class separability measure with different optimization criteria	49
Figure 3-6. Class assignment differences using the separability measure.....	50
Figure 3-7. Optimal classification results based on Bhattacharyya distance.....	51
Figure 3-8. Optimal classification result maps using Bhattacharyya distance	52
Figure 3-9. Class assignment differences using Bhattacharyya distance	53
Figure 4-1. The selected addresses in Volusia County, Florida in 2016	65
Figure 4-2. Reference points to measure positional uncertainty.....	67
Figure 4-3. Histograms of the geocoding positional uncertainties for the different reference points.....	73
Figure 4-4. The empirical cumulative distributions of geocoding positional uncertainty superimposed on their theoretical distribution counterparts	75
Figure 4-5. Boxplots of the Kolmogorov-Smirnov (K-S) statistic values from bootstrapping	77

LIST OF TABLES

Table 2-1. Investigations of visual variables for attribute uncertainty visualization.....	10
Table 2-2. A framework for attribute uncertainty visualization based upon bivariate mapping	15
Table 2-3. Overlaid symbols for uncertainty on a choropleth map.....	16
Table 2-4. Coloring properties for uncertainty on a proportional symbol map.....	17
Table 2-5. Composite symbols for uncertainty visualization.....	18
Table 3-1. Types of cost functions for map classification.....	39
Table 3-2. Configurations for optimal map classification with their acronyms.....	40
Table 3-3. Summary statistics of median household income (in \$) for the 254 counties in Texas	44
Table 3-4. A comparison of optimal classification methods with the class separability and the natural breaks classification methods	46
Table 4-1. Covariates for positional uncertainty in geocoding, together with their abbreviations	70
Table 4-2. Descriptive statistics for positional uncertainty of sample geocoded points	72
Table 4-3. The Kolmogorov-Smirnov (K-S) and the Bayesian information criteria (BIC) diagnostic statistics	76
Table 4-4. Linear regression model results: LM-Building, and LM-Parcel.....	79
Table 4-5. The results of SAR models: SAR-Building and SAR-Parcel	82

CHAPTER 1

INTRODUCTION

The term *uncertainty* is basically used to describe the potential variation in values (Slocum et al. 2009) and is commonly interchangeably used for other terms: error (i.e., bias) and accuracy. However, error, which refers to a deviation from a true value, is difficult to quantify because a true value is generally unknown (Davis and Keller 1997). The term accuracy is a more quantifiable alternative (Buttenfield and Beard 1994), but it also should be carefully used because of its various meanings (Slocum et al. 2009). Specifically, by following recommendations of the Spatial Data Transfer Standard (SDTS), the U.S. Federal Geographic Data Committee (FGDC) lists five fundamental components to address with regard to data accuracy: lineage, positional accuracy, attribute accuracy, logical consistency, and completeness (ANSI 1998). Among these components, this dissertation mainly focuses on positional and attribute accuracy. In addition, this dissertation utilizes the term *uncertainty* because it refers to deviation from a true value without a precise magnitude, and has a narrower definition than the term accuracy (Davis and Keller 1997), which is caused by combining fundamental sources of uncertainty.

Uncertainty inevitably exists in spatial data due to a discrepancy between spatial objects and reality (Longley et al. 2011). Although this discrepancy is augmented through spatial data manipulation and analyses in a Geographic Information Sciences (GIS) environment, it is generally assumed that spatial data are free of uncertainty when they are represented and analyzed (UCGIS 1996). However, visualization of spatial data is highly influenced by uncertainty (Slocum et al. 2009) and could mislead map readers into believing unreliable spatial patterns (Sun, Wong,

and Kronenfeld 2014). From a spatial analysis perspective, uncertainty also increases standard errors and reduces the power of an analysis (Zimmerman and Li 2010).

The purpose of this dissertation is to contribute to extending the understanding of spatial data uncertainty in visualizing and modeling spatial phenomena. Previous studies have developed uncertainty visualization approaches that simultaneously visualize uncertainty along with its corresponding attribute estimates utilizing additional visual variables (e.g., MacEachren et al. 2005). These approaches help to avoid misleading spatial patterns by providing map users with additional reliability information. Furthermore, uncertainty modeling is usually considered as a prerequisite process to reduce uncertainty in spatial data manipulation and analysis (Zhang and Goodchild 2003). Also, the modeling process can help to improve a comprehension of uncertainty because it contains processes for exploring possible covariates of this uncertainty (Zimmerman et al. 2007).

This dissertation investigates three specific sub-topics in visualizing and modeling spatial data uncertainty. First, a framework for attribute uncertainty visualization is suggested based on bivariate mapping techniques, and is implemented in a popular Geographic Information Systems (GIS) environment. A comprehensive framework for uncertainty visualization has been investigated in the literature (e.g., MacEachren 1985; Gahegan and Ehlers 2000; Thomson et al. 2005). However, the framework might not be suitable for an implementation in GIS software due to a lack of detailed strategies. Focusing solely on attribute uncertainty visualization, this dissertation proposes a framework for geovisualization of attribute uncertainty based on the types of attribute and suitable visual variables along with a proper implementation strategy. This framework and implementation utilize a wide range of visual variables that have been investigated

in the literature, and these can provide flexibility to enhance communication and visualization effectiveness. The proposed uncertainty visualization methods are illustrated with census variables at the block group level for the city of Plano, Texas. This research is published in the *Journal of Visual Languages and Computing* (Koo, Chun, and Griffith 2018).

Second, optimal map classification methods for a choropleth map are developed to simultaneously consider attribute estimates and their uncertainty. Although the visualization methods of attribute uncertainty can provide map users reliability information for spatial patterns in a map, the spatial patterns still might be unreliable due to a general map classification that is performed based only on attribute estimates (Sun, Wong, and Kronenfeld 2016). Thus, this dissertation proposes optimal classification methods that utilize the following two dissimilarity measures: the class separability measure (Sun, Wong, and Kronenfeld 2014), and Bhattacharyya distance (Coleman and Andrews 1979). These methods utilize a shortest path solution in an acyclic network based on dissimilarity measures. This research addresses multiple criteria with various cost and objective functions in order to determine an optimal classification result. This research contributes to expanding the discussion about finding an optimal classification result incorporating data uncertainty in a map classification. Furthermore, a graphical diagnostic tool to interactively evaluate and compare optimal classification results has been developed. The proposed optimal classification methods incorporating uncertainty information are demonstrated with census variables for the counties in the state of Texas. It is published in the *Annals of the American Association of Geographers* (Koo, Chun, and Griffith 2017).

Finally, uncertainty modeling for geocoded locations is investigated in this dissertation. The modeling of geocoding positional uncertainty can help to implement imputation methods, and can

allow for realistic positional uncertainty simulations (Zimmerman et al. 2007). In addition, an identification of potential covariates (e.g., properties of street networks) that affect positional uncertainty of geocoded locations can help to detect their impact on spatial analysis results (Zimmerman and Li 2010), and to predict positional uncertainties at specific locations (Jacquez 2012). However, covariates that can provide a comprehensive explanation are still seldom investigated (Zimmerman and Li 2010), and more covariates need to be explored to increase its understanding (Jacquez 2012). Furthermore, spatial autocorrelation among geocoded points has been barely considered in this type of modeling. Thus, this research examines multiple covariates that rarely have been investigated, and considers spatial autocorrelation in regression models. In this research, 22,239 mailable residential addresses in Volusia County, Florida in 2016 are used for the modeling of geocoding positional uncertainty, focusing on positional uncertainty in street geocoding (i.e., address matching). This research is forthcoming in the *International Journal of Applied Geospatial Research* in 2018: volume 9, issue 4.

CHAPTER 2

**GEOVISUALIZING ATTRIBUTE UNCERTAINTY OF INTERVAL AND RATIO
VARIABLES: A FRAMEWORK AND AN IMPLEMENTATION FOR VECTOR DATA***

Authors - Hyeongmo Koo, Yongwan Chun, and Daniel A. Griffith

School of Economic, Political and Policy Sciences

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

NOTE: Koo, H., Y. Chun, and D. A. Griffith. 2018. Geovisualizing attribute uncertainty of interval and ratio variables: A framework and an implementation for vector data. *Journal of Visual Languages & Computing* 44:89–96.

Permission granted republish or display content in this dissertation (4321561091509).

ABSTRACT

Geovisualization of attribute uncertainty helps users to recognize underlying processes of spatial data. However, it still lacks an availability of uncertainty visualization tools in a standard GIS environment. This paper proposes a framework for attribute uncertainty visualization by extending bivariate mapping techniques. Specifically, this framework utilizes two cartographic techniques, choropleth mapping and proportional symbol mapping based on the types of attributes. This framework is implemented as an extension of ArcGIS in which three types of visualization tools are available: overlaid symbols on a choropleth map, coloring properties to a proportional symbol map, and composite symbols.

Key Words: Uncertainty, geovisualization, bivariate mapping, GIS

2.1 Introduction

Spatial data inevitably contain uncertainty in representations of their locations and attributes. The representation and measuring of the world by a data model should have a discrepancy with the real world because the processes of abstracting and generalizing are inevitable attempts to express the real world in a discrete and finite data model (Chrisman 1991). The U.S. Federal Geographic Data Committee (FGDC) lists the following five fundamental components of data quality by adhering to recommendations of the Spatial Data Transfer Standard (SDTS): lineage, positional accuracy, attribute accuracy, logical consistency, and completeness (ANSI 1998). Among these components, location uncertainty, which refers to inaccurate locations of geographic features, often results from global positioning system (GPS) and geocoding errors. Because attributes describing non-locational characteristics of geographic features usually are estimated with samples, attribute uncertainty is caused by sampling errors as well as measurement errors. Furthermore, uncertainty in spatial data can propagate through an aggregation of spatial units, which frequently is necessary due to confidentiality issues, data management issues, computational issues, and representation concerns. Especially, when location and attribute uncertainties simultaneously occur for individual observations, this spatial aggregation process can further increase the level of uncertainty in spatial data (MacEachren 1992). For example, when a sample point is assigned to an incorrect areal unit due to its locational error in an aggregation process (Hay et al. 2009), the location errors further contribute to the attribute uncertainty for an aggregated spatial unit.

Understanding data uncertainty is important and often necessary in various research and practical activities (Deitrick and Wentz 2015). Researchers consider uncertainty that affects data

or a modeling process, such as differences between measured or predicted values and actual or true values. Decision makers usually contend with uncertainty based on how current conditions or policies affect the future results in their decision processes (Erskine et al. 2006). Visualization can provide leverage for a data source to meet social and scientific needs (Bertin 1983). But, a comprehensive visualization of uncertainty is difficult, because it often requires multiple representations of data from different perspectives. Especially, uncertainty visualization commonly requires a simultaneous representation of data in terms of their corresponding attributes on a map. Furthermore, uncertainty visualization can become more complex and more difficult to read and interpret for map readers (Leitner and Buttenfield 2000).

Previous research has investigated visualization methods to effectively and efficiently represent spatial data uncertainty (e.g., MacEachren, Brewer, and Pickle 1998; MacEachren et al. 2005; Sun and Wong 2010). However, uncertainty visualization is still not popular in a geographic information system (GIS) environment. One reason is the unavailability of uncertainty information for attributes (e.g., decennial census), and, hence, the demand for uncertainty visualization in common GIS software is not high, with several exceptions, such as Heuvelink, Brown, and van Loon (2007) and Pebesma, de Jong, and Briggs (2007). The purpose of this paper is to propose a framework for uncertainty visualization methods for vector data, and to prototype implementation of these methods in a standard GIS environment. It focuses on attribute uncertainty visualization of interval and ratio data. Specifically, the coloring properties for proportional symbols (CPPS) and composite symbol (CS) strategies can represent polygon, point, and line features, while the overlaid symbols on a choropleth map (OSCM) strategy is restricted to only polygon features. In

this study, the proposed methods are illustrated with census block group level variables for the city of Plano, Texas using only polygon features.

2.2 Related work

Uncertainty visualization can be considered the same as the display of any other information in conventional cartography (McGranaghan 1993). However, it often requires further attention because it needs to be used for a supplementary purpose to better understand its corresponding attribute (MacEachren 1992). For example, uncertainty information commonly is required to be presented with attribute information in uncertainty visualization. Technically, attribute uncertainty is depicted with visual variables in cartographic representations. Hence investigations in the literature address how Bertin's (Bertin 1983) graphic variables can be logically matched to, and practically utilized for, different uncertainty types. For example, empirical studies (e.g., MacEachren, Brewer, and Pickle 1998; Drecki 2002) investigate the effectiveness of specific visual variables, such as color value, saturation, and opacity. MacEachren (1992) discusses how size, color value, and grain (texture) are useful for depicting uncertainty in numerical information, and how color hue, shape, and orientation can be used for depicting uncertainty in nominal information within the context of Bertin's original graphic variables. Table 2-1 summarizes previous studies with visual variables investigated for attribute uncertainty visualization.

Table 2-1. Investigations of visual variables for attribute uncertainty visualization

Studies	Visual variables
MacEachren (1992)	Color saturation, crispness, transparency, and resolution
Goodchild et al. (1994)	Fog and Texture
Jiang et al. (1995)	Lightness, and saturation
Davis and Keller (1997)	Hue, color value, and texture
Dreki (2002)	Opacity
Hengl (2003)	Saturation and Lightness
Xiao et al. (2007)	Textures
Sun and Wong (2010), Wong and Sun (2013)	Textures
Kubíček and Šašinka (2011)	Lightness and saturation
Slingsby et al. (2011)	Hue and lightness
Kaye et al. (2012)	Saturation
Deitrick and Wentz (2015)	Saturation and size

Color components, including hue, color value (or lightness), and saturation, have been widely used to represent attribute uncertainties among these visual variables. Goodchild, Battenfield, and Wood (1994) and Davis and Keller (1997) suggest that hue, color value, and/or texture can be used for depicting uncertainty. Also, MacEachren (1992) argues that low uncertainty should be represented by pure hues, while high attribute uncertainty should use a less saturated color (i.e., graying out uncertainty features). Jiang, Ormeling, and Kainz (1995) suggest visualization techniques for fuzzy spatial analysis, where lightness and saturation are used to represent uncertainties. Hengl (2003) uses a similar method based on the hue, saturation, and intensity (HSV) color model to depict uncertainty. These two methods rely on both saturation and value to increase whiteness based on the degree of uncertainty. Kubíček and Šašinka (2011) conducted an

intuitiveness testing for Hengl's method, and their empirical test confirmed high uncertainty should be visualized with higher whiteness. More recently, Deitrick and Wentz (2015) merged a visual variable color value with saturation to depict attribute and uncertainty using a bivariate mapping technique.

Transparency and fog also are viable candidates for representing uncertainty information. The metaphor of fog has been proposed by Beard et al. (1991), and an initial application of uncertainty visualization using transparency by MacEachren (1992) was developed based on the metaphor of fog. In this perspective, foggy elements, which make identifying the level of attribute difficult, indicate high uncertainty (MacEachren et al. 2005). Transparency, which also is called opacity, was proposed for integral symbolization. Drecki (2002) asserts that elements with higher transparency (i.e., less opaque objects) are logically considered as more certain things. MacEachren et al. (2012) confirm through empirical tests that transparency is one of the feasible visual variables for uncertainty representation.

Texture has been popularly utilized when attribute uncertainties are represented with classifications. Texture is useful for a binary classification, but also can be used for ordered or numerical data with spacing between hatching in texture symbols (MacEachren 1992). Goodchild, Buttenfield, and Wood (1994) also discuss that texture is a proper visual variable for uncertainty visualization, and can introduce visual noise as a metaphor for noisy data. MacEachren, Brewer, and Pickle (1998) found that texture overlay symbols effectively represent the reliability of death rates, and Xiao, Calder, and Armstrong (2007) illustrate the effect of attribute uncertainty on map classification using texture. Recently, Sun and Wong (2010) and Wong and Sun (2013) also

utilized texture overlay symbols to visualize the quality of American Community Survey (ACS) data.

Statistical maps for uncertainty visualization are constructed in a form of either adjacent display or bivariate mapping (MacEachren 1992). In adjacent display, which also is called a map pair strategy, attribute uncertainty is represented with an additional map accompanying the corresponding attribute map (e.g., MacEachren 1985; Leitner and Battenfield 2000; Aerts, Clarke, and Keuper 2003). The bivariate approach also has been popularly used for uncertainty visualization, because an attribute and its associated uncertainty information may be considered to be a pair of variables. Previous studies conclude that the bivariate approach is more effective than adjacent display in terms of cluster detection (MacEachren, Brewer, and Pickle 1998; Kubíček and Šašínská 2011; Francis et al. 2015). A major advantage of this approach is that readers do not need to switch their focus back and forth between two maps to build connections between the estimates and their uncertainties (Sun and Wong 2010).

2.3 A framework for uncertainty visualization based on bivariate mapping

An attribute and its associated uncertainty generally are visualized using either an adjacent display or a bivariate mapping method (MacEachren 1992). Adjacent display refers to uncertainty being visualized separately from its corresponding attribute, whereas bivariate mapping represents both data and uncertainty in a single map. McGranaghan (1993) argues that the complexity of symbols in a bivariate mapping is beyond a threshold indicting the amount of information decoded at one time. However, uncertainty visualization without interfering with the decoding ability of users has been investigated (e.g., MacEachren 1995; Edwards and Nelson 2001). Furthermore,

studies (MacEachren, Brewer, and Pickle 1998; Edwards and Nelson 2001) show with empirical assessments that bivariate mapping also is more effective than adjacent display. With such support, bivariate mapping for uncertainty visualization primarily has been discussed and implemented as a visual tool in GIS. Implementation of bivariate mapping for uncertainty visualization can be supported by a framework for attribute uncertainty visualization. While previous research, including (Beard et al. 1991; Gahegan and Ehlers 2000; Thomson et al. 2005), attempted to provide a comprehensive framework for uncertainty visualization with considerations of location and attribute uncertainty, research summarized in this paper focuses solely on attribute uncertainty visualization, and suggests a framework for geovisualization of attribute uncertainty based on the types of attribute and visual variables.

When uncertainty visualization based on bivariate mapping is implemented, two issues need to be cautiously decided. First, an appropriate thematic mapping technique needs to be selected to visualize a corresponding attribute considering both the nature of the underlying phenomenon and the purpose of maps (Slocum et al. 2009). Choropleth and proportional symbol maps commonly are used to display data collected for areal units, such as census tracts and counties. A choropleth map is appropriate to display numerical values that are summarized based on polygon boundaries (MacEachren and DiBiase 1991), and to focus on typical values for each individual unit (Slocum et al. 2009). Importantly, data standardization (e.g., population density) often is necessary in choropleth mapping to adjust for varying sizes of areal units. In contrast to the choropleth map, a proportional symbol map is used to represent raw counts or totals (e.g., number of people). A proportional symbol map also can be complementary to a choropleth map. Hence, when attributes

are visualized in a thematic map with uncertainty, different thematic mapping techniques should be used, depending on the types of attributes.

Second, uncertainty visualization with a bivariate mapping approach can be categorized into either a visually integral method or a visually separable method (MacEachren, Brewer, and Pickle 1998). The visually integral method modifies one visual variable to represent both attribute and uncertainty. For example, color can be modified with hue and saturation; hue usually represents an attribute, and saturation is suitable to represent uncertainty. The visually separable method adds another visual variable to represent uncertainty. For instance, usually hue and lightness of color are used to represent an attribute, and an additional visual variable, such as spacing between hatching in overlaid texture symbols, is combined to describe uncertainty. Researchers (MacEachren, Brewer, and Pickle 1998; Slocum et al. 2003) comment that the visually separable method is better than the visually integral method in terms of extracting information. Thus, uncertainty should be represented by other visual variables or symbols when bivariate mapping is employed for uncertainty visualization.

Table 2-2 presents a framework for uncertainty visualization using a bivariate mapping technique that is proposed based upon the two preceding considerations. In this framework, standardized attributes are visualized with choropleth mapping, and their uncertainty is represented with overlaid symbols, such as texture and bars. Because hue and lightness typically are utilized for portraying an attribute in a choropleth map, other variables, including spacing, height, and size, can serve to effectively represent uncertainty. When a proportional symbol map is used for an unstandardized attribute, visually separable variables (e.g., color value, saturation, and size) can be used to display attribute uncertainty.

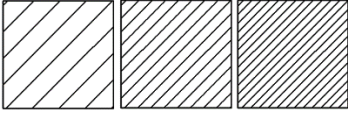
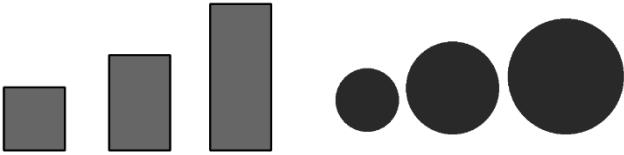
Table 2-2. A framework for attribute uncertainty visualization based upon bivariate mapping

Type of Attribute	Thematic mapping techniques	Additional visual variable for uncertainty representation
Standardized data	Choropleth mapping	Spacing, sizes, height
Unstandardized data	Proportional symbol maps	Saturation, color value, size

2.4 An implementation of uncertainty visualization


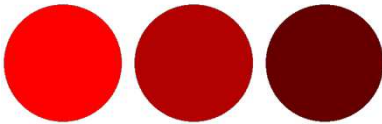
Attribute uncertainty visualization based on bivariate mapping is implemented in ArcGIS in accordance with the preceding framework. First, an OSCM method is implemented to visualize attribute uncertainty. A choropleth map is suitable to represent attributes at the ratio scale, and the colors of the symbols (usually hue and lightness) in the choropleth map are used to represent uncertainty values. Additional overlaid symbols to visualize attribute uncertainty include textures (spacing), circles (size), and bars (size). A texture symbol is one of the most widely used visual variables to represent uncertainty in previous results (e.g., Goodchild, Battenfield, and Wood 1994; Xiao, Calder, and Armstrong 2007; Wong and Sun 2013). Texture symbols are the best in a binary classification (MacEachren 1992), but also can be employed for numerical variables using spacing between hatching (Sun and Wong 2010). Symbol sizes also can be utilized for depicting uncertainty (e.g., Deitrick and Wentz 2015), although they can be confusing when appearing with attribute values (Slocum et al. 2009). Table 2-3 illustrates uncertainty mapping with these three overlaid symbol maps.

Table 2-3. Overlaid symbols for uncertainty on a choropleth map

Attribute	Uncertainty (Low ↔ High)	
	Spacing	Size
Choropleth map		

Second, CPPS is implemented to represent uncertainty in a GIS environment. A proportional symbol map is more appropriate to represent raw counts or frequencies (Slocum et al. 2009). The graphical sizes of symbols can be defined by a mathematical scaling based on the minimum and maximum values of attributes. Attribute uncertainty can be represented with color saturation or color value in HSV color model of proportional symbols. Table 2-4 graphically illustrates this approach. The saturation and color value linearly decrease from those of an original color to its minimum value (for which zero or a predefined quantity can be used). Thus, in terms of saturation, an attribute value with a high level of certainty is represented with pure hues, whereas those with low levels of certainty are visualized with less saturated colors (i.e., graying out uncertainty features) (MacEachren 1992). With regard to color value, highly certain attribute values are represented with pure hues, and, in contrast, highly uncertain attribute values are represented with low color values (e.g., black) (Jiang, Ormeling, and Kainz 1995; Davis and Keller 1997).

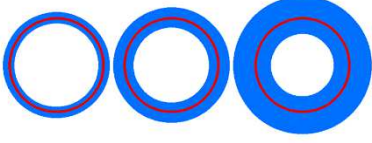
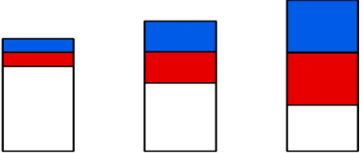
Table 2-4. Coloring properties for uncertainty on a proportional symbol map

Attribute	Uncertainty (Low ↔ High)	
	Saturation	Color value
Size (proportional symbols)		

Standard errors, which measure a certain type of attribute uncertainty, enable visualizing a range of attribute values with a given degree of confidence, based upon a specific probability model. This confidence interval can be depicted with an individual symbol. Table 2-5 illustrates composite symbols in which both attribute values and such corresponding uncertainties are visualized. Conceptually, these composite symbols are constructed with three different proportional symbols overlaid for each individual location. Two of these three proportional symbols represent uncertainty by visualizing the upper and lower limits of attribute values at a given confidence level. Accordingly, this range, which is represented by the thickness of the circle symbols in Table 2-5, portrays a possible range of an attribute value. Corresponding attribute values appear in the range constructed by the upper and lower bound lines (e.g., the center line in the circle symbols in Table 2-5). The sizes of three symbols can be proportionally defined with corresponding values. With bar chart symbols, first and second items on the top of the symbol represent the lower and upper confidence limits of an attribute value. The common boundary of lower and upper confidence areas represents an attribute value. These two approaches using a composite symbol allow users to directly compare uncertainties with corresponding attribute values and their confidence intervals. Moreover, with the circle composite symbols, a thicker bound around the center line indicates a wider confidence bound; in other words, a thick symbol

indicates that its standard error or attribute uncertainty is large. For a bar chart composite symbol, the level of uncertainty is represented by the combined heights of the first and second items.

Table 2-5. Composite symbols for uncertainty visualization

Attribute	Uncertainty (Low ↔ High)	
	Size of circles	Height of charts
Size (proportional symbol or chart)		

The proposed methods have been developed as an extension in ArcGIS 10.1 using C# on the Microsoft .Net Framework 4.¹ This geovisualization extension of attribute uncertainty provides three visualization functions: OSCM, coloring properties to a proportional symbol map, and composite symbols. Figure 2-1 portrays the menu interface of the extension. Each function has its own graphic user interface (GUI), and the functions allow users to select fields for an attribute and its uncertainties with proper visual variables. The GUIs are presented in the Appendix A.

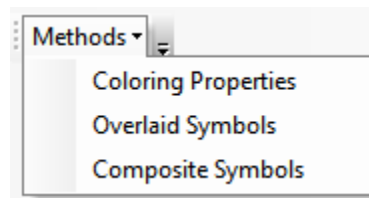


Figure 2-1. Menu interface of the attribute uncertainty geovisualization extension

¹ <https://hdl.handle.net/10735.1/5609>

2.5 An application

The application presented in this section illustrates the implemented methods in ArcGIS using the 5-year ACS data (2009-2013). For the interval/ratio measurement scale, by definition the standard error is the standard deviation based upon sampling error. This quantifies the attribute uncertainty due to sampling (Sun and Wong 2010). These data include a margin of error (MOE) that is based upon the standard error of the sampling distribution for an attribute variable, as well as survey estimates. Because MOEs represent the lower and upper bounds for a 90-percent confidence level, the standard error of an estimate can be calculated directly from its MOE: that is, $MOE = 1.645 \times \text{standard error}$ (U.S. Census Bureau 2009). However, some studies (Li and Zhao 2005; Sun and Wong 2010) discuss that a visual representation of absolute error measures (e.g., the standard error of an estimate) may lead to an inappropriate conclusion and interpretation because large attribute values tend to have large standard errors. Furthermore, Francis et al. (2015) discuss the appropriateness of using relative measures of errors, such as the coefficient of variation (CV), for counts (e.g., the total number of households and of occupied houses) and median or mean values (e.g., median age and mean household income). Nevertheless, an absolute measure of error is still appropriate for proportional or ratio values (e.g., percentage of Asians), which are bounded by a specific value range: i.e., between 0 and 1 for proportional values, and between 0 and 100 for percentages (Francis et al. 2015) .

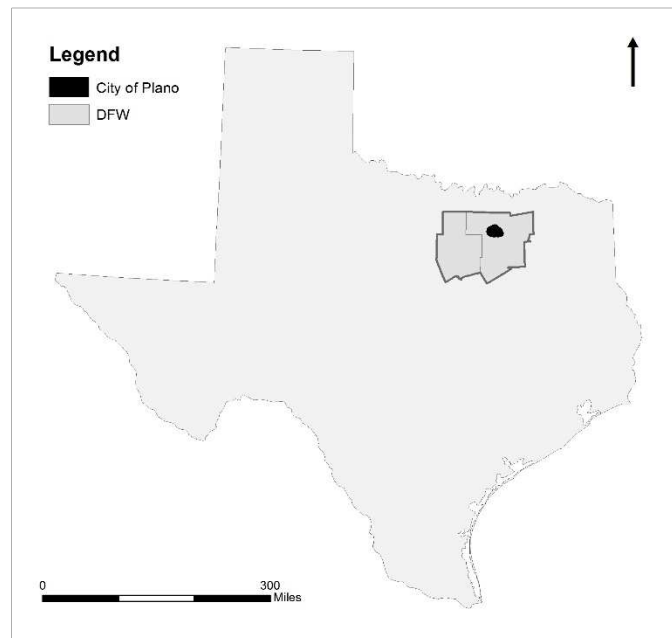


Figure 2-2. Study Area

Two variables with their corresponding MOEs were obtained from the 5-year ACS data (2009-2013) at the census block group level for the city of Plano, Texas, which is located in the northeastern part of the Dallas-Fort Worth metropolitan area (Figure 2-2). Median age was chosen to illustrate a choropleth map with overlaid symbols, and ‘the number of housing units’ was chosen to illustrate proportional symbols with coloring properties and composite symbols. A relative measure of errors, the CV, is used to represent attribute uncertainty for both variables.

2.5.1 Overlaid Symbols on a Choropleth Map (OSCM)

Figure 2-3 portrays visual representations of 2013 median age with its CV at the census block group level for the city of Plano using the OSCM method. Color symbols (specifically, hue and brightness) of the choropleth map represent the estimated median ages. The estimates are classified

with four categories and the natural break algorithm. Attribute uncertainties are visualized with three different types of overlaid symbols; textures, bars, and circles. Spacing between hatching in the texture overlay, heights of the bar symbols, and circle sizes furnish a way to represent the degrees of the CV of median age estimates; thus small spacing in texture, tall bars, and large circles describe a high level of uncertainty in estimated median age values. Because median age estimates and their uncertainties are integrated into one display frame, map readers need to concentrate on only one frame. Thus, users can easily build connections between the estimates and their corresponding CV levels. In the median age estimate maps, census block groups in the southern part of the city have relatively higher median age estimates. However, mapping only the estimates may not provide sufficient information with which map readers can easily recognize the credibility of the differences across census block groups. Census block groups with relatively higher CV levels also are observed in the southern part of the city, which indicates that the census block groups in this region contain a higher level of uncertainty in terms of their CV levels.

However, these maps show some potential disadvantages of this visualization method. First, determining corresponding classes from texture overlay symbols is difficult, especially when small areal units have only a small number of lines because of large line spacing separations. There is no general guide to determine the largest spacing value, but each areal unit should be displayed with at least two lines to utilize spacing for a texture symbol. Hence, the largest spacing value should be judiciously chosen according to sizes of areal units and their corresponding attribute uncertainty levels. Second, the large size of overlaid circles or the tall height of overlaid bars can partially cover color on a choropleth map, and can cause problems in interpreting the distribution of estimates. The sizes of overlaid symbols may need to be restricted so that the underlying

geographic distribution of estimates can be effectively displayed. In most cases, attributes are still more important than their accompanying uncertainties.

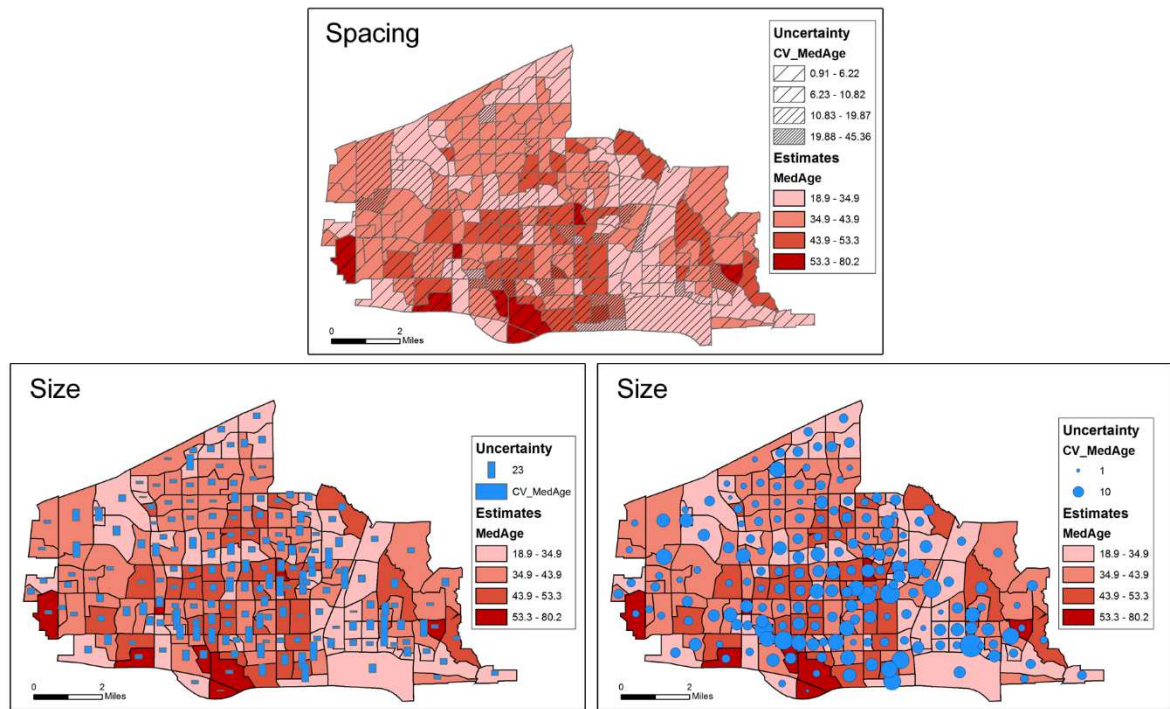


Figure 2-3. OSCM for 2013 median age and its coefficient of variation by census block group in Plano, TX

2.5.2 Coloring Properties on a Proportional Symbol map (CPPS)

Figure 2-4 visualizes the estimated numbers of housing units and their uncertainty at the census block group level in the city of Plano. The sizes of the circle symbols represent the numbers of housing unit estimates, and their uncertainties are portrayed with saturation and color value in terms of the HSV color model. The census block groups with low saturation and color values represent high uncertainty for the number of housing unit estimates. With a visually separable bivariate mapping technique adopted for this visualization tool, the additional visual variables

(saturation and color value) are combined into proportional symbols. Hence, this map should help map readers to easily build linkages between estimates and their corresponding CVs, as MacEachren, Brewer, and Pickle (1998) claim.

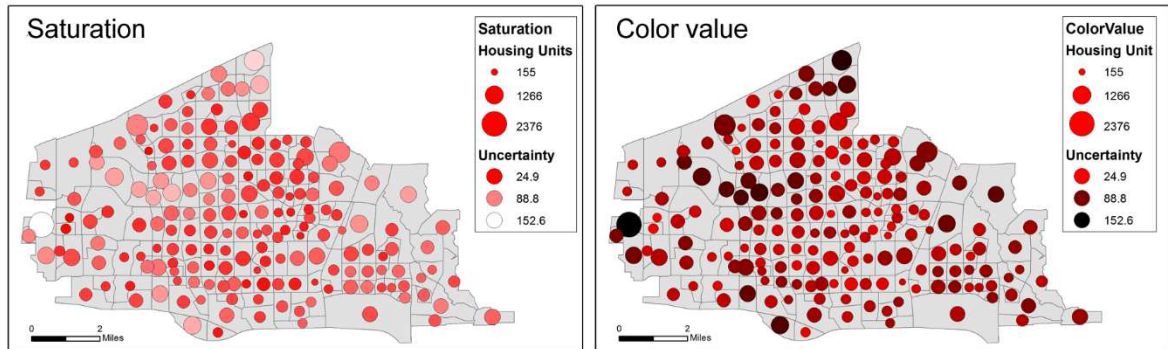


Figure 2-4. Coloring properties on proportional symbol maps for the 2013 number of housing units and its coefficient of variation by census block group in Plano, TX

This map also has some disadvantages that stem from common problems of proportional mapping. When the difference between the maximum and minimum values is extremely large, the map pattern can be difficult to read due to massive overlaps of symbols (Slocum et al. 2009). Although the drawing order of the symbols is controlled with regard to the symbol size range (i.e., the largest symbol to the smallest symbol) in order to prevent small symbols from being concealed by larger symbols, the minimum symbol size needs to be selected carefully in order to achieve communicative and visual effectiveness of a map. In addition, the legend for such maps may become complicated. Such a legend has two legend groups. Each of them shows, respectively, the distribution of estimates and uncertainties (CVs) with their own minimum and maximum values. But, because the CVs show only relative uncertainty for estimates, the legend does not necessarily clarify actual differences between estimates.

2.5.3 Composite Symbols (CS)

Figure 2-5 portrays composite symbol maps for the number of housing unit estimates with 99 percent confidence intervals based on a normal distribution assumption. In the top map, the center lines in the proportional symbols represent the estimates, and the bounds around the center lines portray corresponding confidence intervals. A thick bound indicates high uncertainty of an attribute value. For the bar chart symbol map, the two areas on the top of stacked bars respectively represent the upper and lower bounds of the attribute values. The common boundary between the first and second items on the top of a symbol in the symbols represents attribute values. That is, the number of housing unit estimates is represented by the combined height of the second and third items. The level of uncertainty is represented by the combined heights of the first and second items. The variable used in the composite symbols method in Figure 2-5 is identical to the one in the map displayed with CPPS in Figure 2-4. Although a relative comparison is only possible in the map using CPPS, this composite symbol map allows map readers to compare actual differences of estimates for a given confidence level, based on the sizes of the symbols. Composite symbol maps are particularly useful for simultaneously identifying uncertainty information in specific locations and general spatial patterns of an attribute. When a spatial pattern of an attribute is emphasized, an increase in symbol sizes can enhance the spatial pattern. When a user focuses on local data exploration, symbols can be adjusted accordingly. The implemented tool furnishes an easy way to change symbols as well as exploration functions, including zoom in/out. However, a composite symbol map has several potential issues that originate with proportional symbol mapping. These include an excessive number of overlaps and legibility of small size symbols (Slocum et al. 2009). Distinguishing the border between the estimate and confidence bounds is difficult especially when

symbol size is small and/or uncertainty is relatively small for a corresponding estimate. Also the level of legibility can vary depending on a user's familiarity with symbols used.

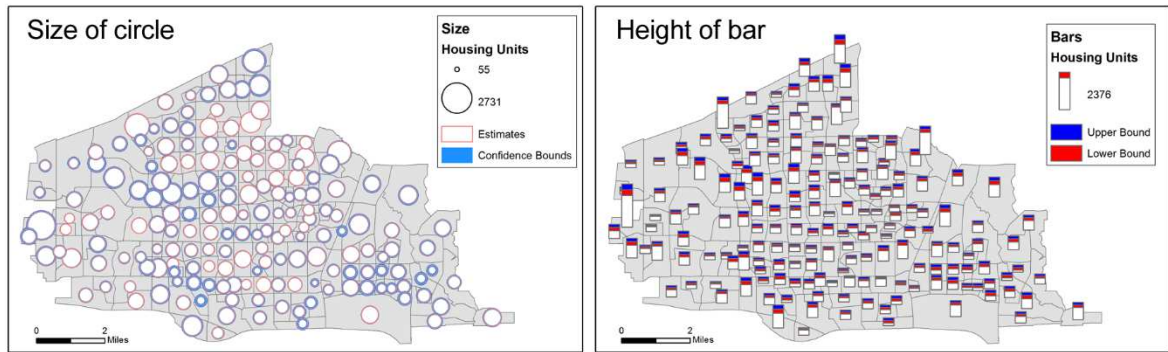


Figure 2-5. Composite symbol maps for the 2013 number of housing units with 99 percent confidence intervals by census block group in Plano, TX

2.6 Summary and conclusions

Geovisualization of spatial data can be enhanced by providing a visual representation of their attribute uncertainty, which in turn can assist users to more easily recognize an underlying pattern of data. However, it still lacks such tools in a GIS environment, and a general framework for uncertainty visualization. This paper proposes a framework for attribute uncertainty visualization based on bivariate mapping techniques, and provides such tools implemented in the ArcGIS environment. This framework principally utilizes choropleth maps and proportional symbol maps for thematic mapping of attributes based on the types of variables: count and ratio estimates. This implementation contains three types of visualization functionalities: overlaid symbols on a choropleth map, coloring properties for a proportional symbol map, and composite symbols. Many

of the visual variables that have been investigated in the literature are available to provide flexibility to enhance communication and visualization effectiveness.

In this paper, the visualization tools were examined with a single dataset comprising 84 areal units. In addition, these visualization methods may need to be fine-tuned for specific datasets. For example, when a proportional symbol map is used, a large portion of symbols can overlap with each other, and, for another example, texture symbols with hatchings in a choropleth map may not be easily recognized for small areal units. In these circumstances, minimum and/or maximum symbol sizes need to be carefully set. Similarly, the maximum spacing between hatch lines also needs to be wisely set. Although a composite symbol map can allow map readers to directly compare differences between estimates and their uncertainties, small differences still are difficult to identify. Furthermore, a focused group survey conducted as future research would be useful to evaluate the effectiveness of the implemented tools with various geographical scales and resolutions.

Acknowledgement

This research is supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

CHAPTER 3
OPTIMAL MAP CLASSIFICATION INCORPORATING UNCERTAINTY
INFORMATION *

Authors - Hyeongmo Koo, Yongwan Chun, and Daniel A. Griffith

School of Economic, Political and Policy Sciences

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

NOTE: Koo, H., Y. Chun, and D. A. Griffith. 2017. Optimal map classification incorporating uncertainty information. *Annals of the American Association of Geographers* 107 (3):575–590. Permission granted republish or display content in this dissertation.

ABSTRACT

A choropleth map frequently is used to portray the spatial pattern of attributes, and its mapping result heavily relies on map classification. Uncertainty in an attribute has an influence on map classification, and, accordingly, can generate an unreliable spatial pattern. However, only a few studies have explored the implications of uncertainty in map classification. Recent studies present methods to incorporate uncertainty in map classification and generate a more reliable spatial pattern. Nevertheless, these methods often produce an undesirable result, with most observations assigned to one class, and struggle to find an optimal result. The purpose of this paper is to expand the discussion about finding an optimal classification result considering data uncertainty in a map classification. Specifically, this paper proposes optimal classification methods based on a shortest path problem in an acyclic network. These methods utilize dissimilarity measures and various cost and objective functions that simultaneously can consider attribute estimates and their uncertainty. Implementation of the proposed methods is in an ArcGIS environment with interactive graphic tools, illustrated with a mapping application of the American Community Survey data in Texas. The proposed methods successfully produce map classification results achieving improved homogeneity within a class.

Key Words: uncertainty, map classification, choropleth map, GIS

3.1 Introduction

Uncertainty is unavoidably included in spatial data due to discrepancies (usually due to measurement error) between the real world and observations (Longley et al. 2011). The United States (U.S.) Federal Geographic Data Committee (FGDC) describes spatial data uncertainty in terms of five fundamental components of data quality: lineage, positional accuracy, attribute accuracy, logical consistency, and completeness (ANSI, 1998). Among these components, measurement and sampling errors are the main causes of attribute accuracy or uncertainty, and increasingly uncertainty information is being reported together with estimates from survey data, such as margins of error (MOE) for the American Community Survey (ACS) data. Attributes often are visualized with thematic mapping techniques, and attributes measured with an interval or ratio scale are visualized with a choropleth map. Map classification, which has a substantial impact on the spatial pattern of a geographical phenomenon portrayed by a choropleth map, also potentially is influenced by data quality [i.e., uncertainty (Slocum et al. 2009, 425)]. Furthermore, uncertainty can provide map users with insight into the reliability of spatial patterns. However, uncertainty is largely ignored in choropleth mapping; rather, the presumption seems to exist that spatial data and maps are accurate (Sun, Wong, and Kronenfeld 2017).

Incorporating uncertainty in mapping has been investigated in two major ways. First, many studies have developed approaches that simultaneously visualize uncertainty and its corresponding attribute estimates (i.e., observed values) using additional visual variables (e.g., MacEachren et al. 2005). These visualization methods can provide users with reliability information for spatial patterns in a map, which still might be unreliable (Sun, Wong, and Kronenfeld 2017). That is, uncertainty is provided in a separate form than attribute estimates, but map classification is

performed based only on estimates. Second, several recent studies propose methods to incorporate uncertainty information within the map classification process. They focus on reliability of map classification results considering the uncertainty of attribute estimates. Sun, Wong, and Kronenfeld (2014, 2016) accomplish this goal by utilizing a statistical significance of differences between two observations based on their estimated values and their uncertainties. These researchers find class breaks that maximize statistical differences between classes, assuming that the distribution for each observation follows a bell-shaped curve in terms of its estimate and uncertainty. However, this method's map classification is sensitive to outliers because an outlier almost always is significantly far from other observations. As a result, this method often produces a class with only a single outlier, assigning most observations to one (or two) classes. Sun, Wong, and Kronenfeld (2016) furnish a heuristic method utilizing multi-criteria for grouping purposes, which include a class separation level, within class variation, and balance in the number of observations across classes. This approach provides a tool to help a user choose a set of threshold values for those characteristics, but finding an optimal classification remains difficult.

This study expands the notion of incorporating data uncertainty into optimal classification methods by utilizing the following two similarity measures that simultaneously consider attribute estimates and their uncertainty: the class separability measure (Sun, Wong, and Kronenfeld 2014) and Bhattacharyya distance (Coleman and Andrews 1979). The research summarized here also addresses multiple criteria to determine an optimal classification result. Furthermore, a graphic diagnostic tool to interactively evaluate and compare optimal classification results has been developed and is presented. The proposed optimal classification methods incorporating uncertainty information are illustrated with county-level variables for the state of Texas.

3.2 Literature Review

Spatial data uncertainty has been investigated in relation to GIScience and remote sensing (Goodchild and Gopal 1989; Foody and Atkinson 2002; Zhang and Goodchild 2002; Couclelis 2003; Chiles and Delfiner 2009). A number of studies have been conducted to: identify sources of uncertainty (e.g., Chrisman 1991), detect and measure uncertainty in spatial data and analysis results (e.g., Cayo and Talbot 2003; Bichler and Balchak 2007; Griffith et al. 2007), model or predict analysis results associated with uncertainty (e.g., Zimmerman et al. 2007; Zandbergen and Hart 2009), and develop methods or strategies to reduce or eliminate uncertainty in data and their analysis (e.g., Zimmerman 2008; Goldberg 2011). Also, uncertainty visualization now constitutes a sizeable research domain, and people in cartography, geography, and computer science have contributed to its development (e.g., MacEachren 1992; Pang 2001; Aerts, Clarke, and Keuper 2003; Johnson and Sanderson 2003).

Uncertainty visualization is an essential tool to help map users understand the reliability of spatial patterns for various research and practical activities (Deitrick and Wentz 2015). However, the visualization of uncertainty has been challenging, because it often requires displaying more information than conventional cartographical techniques can handle effectively in a single map. That is, a concurrent representation of data and their corresponding uncertainty information becomes complex and difficult to read and interpret (McGranaghan 1983). Thus, studies have investigated effective and efficient visualization methods for spatial data uncertainty (e.g., Goodchild, Buttenfield, and Wood 1994; MacEachren, Brewer, and Pickle 1998; MacEachren et al. 2005). Notably, using empirical subject tests, Leitner and Buttenfield (2000) report that inclusion of uncertainty information provides clarification rather than an increase in complexity.

Because uncertainty information should be presented with its corresponding attributes, many previous studies addressed uncertainty visualization by utilizing additional visual variables [see Bertin, (1983), for a comprehensive discussion of visual variables]. Among these visual variables, color components, such as color value, lightness, and saturation, have been widely used to describe uncertainty (e.g., MacEachren 1992; Jiang, Ormeling, and Kainz 1995; Leitner and Buttenfield 2000; Hengl 2003; Deitrick and Wentz 2015). Also, transparency and fog (e.g., Beard et al. 1991; Drecki 2002; MacEachren et al. 2005), and texture (e.g., Goodchild, Buttenfield, and Wood 1994; MacEachren, Brewer, and Pickle 1998; Xiao, Calder, and Armstrong 2007; Sun and Wong 2010; Wong and Sun 2013) are practical candidates for describing uncertainty. These visualization methods can be used effectively to warn map readers about the reliability of spatial patterns, but the spatial patterns generated only with estimates can remain unreliable (Sun, Wong, and Kronenfeld 2017). For example, although they are statistically different, observations can be allocated to the same class. Recent map classification methods incorporating uncertainty information suggest a need to improve the reliability of class breaks.

Optimal map classification has been investigated extensively in thematic mapping. Jenks's natural breaks method, which is equivalent to unconstrained clustering (Fisher 1958), is popular and is utilized in extensively GIS packages (Dent 1999). Fundamentally, because Jenks's method is a within group variance minimization approach, it determines class breaks to maximize homogeneity within classes (Murray and Shyy 2000). Cromley (1996) pursued related work. His approach considers map classification as a one-dimensional partitioning problem that also maximizes within group homogeneity. Cromley's work has been further extended to a bi-criterion (Murray and Shyy 2000) as well as a multi-criteria optimization framework (Armstrong, Xiao, and

Bennett 2003). However, implications of data uncertainties have not been explored in this optimization framework.

Xiao, Calder, and Armstrong (2007) propose a robustness measure to explicitly evaluate the effects of data uncertainty in map classification. Their robustness measure is formulated with a probability that an attribute estimate falls into the class where the value is assigned. They also define a tolerance level for the unreliability of a map classification result, which can be interpreted as a minimum acceptable level for map users. Then, the percentage of observations within a given tolerance level is used to quantify overall classification robustness. They utilize this robustness measure to evaluate classification results based on different classification methods (e.g., equal interval, quantile, and natural breaks). This robustness measure is useful for evaluating the overall performance of a map classification result, and provides a warning to map users about the unreliability of class breaks (Sun, Wong, and Kronenfeld 2017). Sun and Wong (2010) suggest a modified natural breaks algorithm that extends the robustness measure. This algorithm identifies breaks using the significance of statistical differences between ordered attribute estimates. In detail, after sorting observations in ascending order based on their estimates, a statistical difference test is conducted between all pairs of consecutive observations. If a test result is significant, a class break is inserted between observations. One limitation of this approach is that a statistical test is conducted only between pairs of consecutive observations. The minimum probability value for this test can be for nonconsecutive observations. Another limitation is that the number of classes depends on a preset significant level, and, accordingly, map classification can result in too few or too many classes.

Sun, Wong, and Kronenfeld (2014) propose a class separability classification method by extending Sun and Wong's modified natural breaks procedure. This class separability classification method uses the standard two-sample z-test² to measure separability between two observations, with their estimates as means and their uncertainties as (unequal) variances. In most empirical cases, these observations are sampled data accompanied by a sample mean and sample variance (e.g., ACS data). A class separability measure between two classes is defined with the minimum confidence levels of the standard two-sample z-test among all pairwise combinations of observations in the two classes. Then, class breaks are determined with k (a preset number of classes) greatest class separabilities, or a preset significant level for the class separability measure. Consequently, class breaks are determined between two ordered estimates, with little overlap in their probability density functions that represent spreads of estimates based on their uncertainties. Although it is simple and fast, one weakness of this approach is that classification results from this method are sensitive to outliers. Hence, this method often produces a classification result in which most observations are assigned to a small number of classes, which is less desirable to display an underlying spatial pattern and to achieve a visual balance. Sun, Wong, and Kronenfeld (2016) put forward a heuristic classification approach that considers not only the class separability measure, but also other classification criteria, such as the unevenness of observation counts across classes, and within-class variability. The consideration of additional criteria can improve a map classification result in terms of balance among classes. For example, the degree of unevenness, which can be calculated as the standard deviation of the number of observations by class, can help

² The standard two-sample z-test compares two means whose sampling distributions conform to a normal distribution with known standard deviations.

achieve a balanced result in terms of the number of observations for each class. Also, within-class variability, one criterion utilized in Jenks's natural breaks method, can ensure that observations with similar estimates are assigned to the same class (Slocum et al. 2009). This heuristic method allows map users to find class breaks with reasonably high class separability measures, while satisfying other map classification criteria. However, untrained users still might encounter difficulties finding appropriate values for criteria that yield an optimal map classification.

3.3 Method

This section summarizes the formulation of an optimal map classification that utilizes a shortest path solution for an acyclic network (Cromley and Campbell 1991). While costs are defined as distance in a shortest path problem, they can be defined as a form of similarity or dissimilarity among observations within a class in a map classification problem. Here, a class separability measure and Bhattacharya distance are used to calculate costs taking into account uncertainty and attribute estimates.

3.3.1 Optimal Classification

An optimal classification problem, whose objective is to maximize internal homogeneity, can be formulated as a p -median problem for facility locations, whose objective is to minimize a total travel weighted distance, when the rank ordered attribute values are represented as points on a simple network (Monmonier 1973). A network structure for classification problems is much simpler than a network structure for p -median problems, because the network for a classification

problem is acyclic (Cromley and Campbell 1991). This section summarizes optimal classification methods that incorporate uncertainty information, extending the shortest path problem of Cromley (1996) to classify observations in an acyclic network. Figure 3-1 shows an example of an acyclic network with five nodes. Let N be the set of nodes in an acyclic network. In addition to starting and ending points of a network, the nodes that correspond to class breaks are located between observations; thus, the number of nodes in N is $n + 1$ for n observations. The lines denote ordered pairs (i, j) , where $i \in N$, $j \in N$, and $i < j$, which represent the different groupings that include observations between nodes i to j . Given n observations in a dataset, there are $n(n + 1)/2$ possible lines in an acyclic network. A cost value associated with a line represents the cost of traversing that line, and, hence, this cost commonly is measured with a type of physical distance between two nodes in a shortest path problem. In a map classification problem, a cost value often is measured with a level of heterogeneity (or homogeneity) among observations in the same class. For example, in univariate map classification without considering uncertainty, a cost value can be calculated as the sum of deviations from a class mean. Then, an optimal solution can be found with a path that minimizes a total cost (or maximum cost) subject to a preset number of lines, which corresponds to the number of classes. In Figure 3-1, thicker lines of $(0, 2)$ and $(2, 4)$ highlight a sample solution with two classes, which respectively represent one class with observations between nodes $(0, 2)$, and another class with observations between nodes $(2, 4)$.

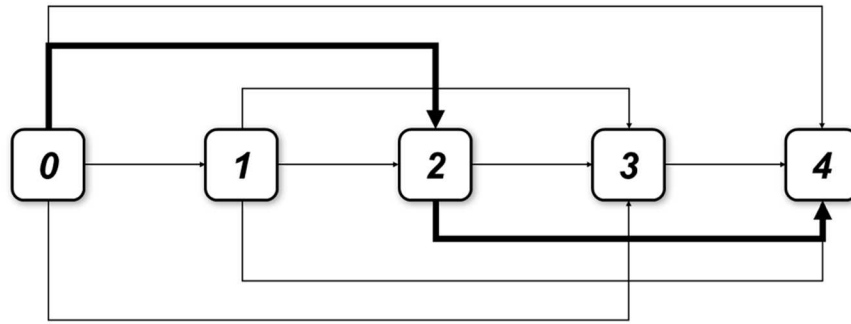


Figure 3-1. An example of an acyclic network with five nodes: the thicker lines of (0, 2) and (2, 4) represent a sample solution with two classes [one class with observations between nodes (0, 2), and another class with observations between nodes (2, 4)]

3.3.2 Cost Functions

An optimal classification result is strongly affected by the definition of a cost function (Cromley 1996). In univariate classification without considering uncertainty, costs commonly are measured with the sum of deviations from a central tendency measure for each class, such as a class mean (Jenks 1977). However, no obvious central tendency measure exists when observations are compared with their uncertainty as well as their estimates, and, subsequently, deviations from a selected central tendency measure are not easily defined. This situation suggests a need to use alternative measures for heterogeneity among observations in the same class.

In the research supporting this section, a cost value (i.e., heterogeneity within a class) was computed as a composite value, accounting for uncertainty, with all pairwise distances among observations in the class. The following two measures were employed to calculate the distance (i.e., the difference) between two observations: a class separability measure, and Bhattacharyya distance. First, a class separability measure (Sun, Wong, and Kronenfeld 2014) was used to calculate differences between observations using their attribute estimates with uncertainties. This

class separability utilizes a statistical difference assuming that observations follow a normal distribution, because a distribution of uncertainty converges to a normal distribution based upon the Central Limit Theorem. This class separability measure can be expressed as

$$Z_s(i, j) = \frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{SE_i^2 + SE_j^2}}, \quad (1)$$

where $|\bar{x}_i - \bar{x}_j|$ is the absolute difference between two estimates, and SE_i and SE_j are their respective standard errors. This equation is derived from the standard z-test, and z-scores are used to quantify the distance (or the difference) between two observations.

The second measure is Bhattacharyya distance, which is a dissimilarity measure that quantifies the dissimilarity of two discrete or continuous probability distributions (Bhalerao and Rajpoot 2003). Dissimilarity measures have been widely used in clustering methods (Mennis and Guo 2009), such as K-means, and self-organizing maps (SOMs) (Kohonen 2001). Bhattacharyya distance frequently is used in feature selection and extraction for remotely sensed images (e.g., Schmidt and Skidmore 2003; Mas et al. 2004). Assuming that an observation follows a known distribution (e.g., normal), Bhattacharyya distance can be utilized to quantify differences between observations accounting for both their attribute estimates and uncertainties. Bhattacharyya distance between two normal distributions can be calculated with the following equation (Coleman and Andrews 1979):

$$D_B(i, j) = \frac{1}{4} \ln \left(\frac{1}{4} \left(\frac{SE_i^2}{SE_j^2} + \frac{SE_j^2}{SE_i^2} + 2 \right) \right) + \frac{1}{4} \left(\frac{(\bar{x}_i - \bar{x}_j)^2}{SE_i^2 + SE_j^2} \right), \quad (2)$$

where \bar{x}_i and \bar{x}_j are the attribute estimates for observations i and j , and \ln denotes the natural logarithm. The first term quantifies the difference between the variances, and the second term compares both estimates and variances. Thus, Bhattacharyya distance can compute a non-zero

distance between two observations with the same estimate and different variances. In contrast, the class separability measure always returns zero for such distributions.

A cost is calculated, based on distances between pairs of observations, in two different ways. First, the sum of pairwise distances in a class is utilized. A large sum of pairwise distances can indicate that some observations in a class tend to differ from other observations in the same class. Second, the maximum pairwise distance in a class is used for its cost value. This quantity controls for the worst case scenario when calculating homogeneity within a class. Table 3-1 presents combinations of the two pairwise distance measures, and two ways to compute cost values. In Table 3-1, $G_{i,j}$ represents all pairs of observations within a grouping.

Table 3-1. Types of cost functions for map classification

Cost determination	Measures	
	The class separability	Bhattacharyya distance
Sum of pairwise distances	$c_{i,j} = \sum_{k \in G_{i,j}} Z_s(k)$	$c_{i,j} = \sum_{k \in G_{i,j}} D_B(k)$
Maximum pairwise distance	$c_{i,j} = \max_{k \in G_{i,j}} Z_s(k)$	$c_{i,j} = \max_{k \in G_{i,j}} D_B(k)$

3.3.3 Objective functions and formulations

A shortest path problem can be solved using either the Dijkstra (1959) or other heuristic algorithms, such as A* (Hart, Nilsson, and Raphael 1968). Also, integer programming (IP) is a well-known alternative strategy. IP problems where the objective function and all constraints are linear can be solved relatively effectively using a branch-and-bound algorithm (Dean 2011). A simple IP problem can furnish a way to solve the proposed map classification problem. This paper

presents an optimal map classification formulation with two different objective functions based on the cost functions described in the previous section. The first objective is to minimize a total sum of costs, focusing on the overall efficiency of a map classification. The second is to minimize the maximum cost among classes, which improves efficiency by controlling for the worst case. Fundamentally, it utilizes the Minimax principle, which commonly is applied to a facility location problem (e.g., Toregas and ReVelle 1972; Erkut, Francis, and Tamir 1992). Accordingly, Table 3-2 shows that eight configurations for optimal classification models relate to the two objective functions, two cost determination methods, and two pairwise distance measures.

Table 3-2. Configurations for optimal map classification with their acronyms

Objective functions	Cost determination	Measures	
		The class <u>S</u> eparability	<u>B</u> hattacharyya distance
Minimize	Sum of pairwise distances	SSS	SSB
a total <u>S</u> um of costs	<u>M</u> aximum pairwise distance	SMS	SMB
Minimize	Sum of pairwise distances	MSS	MSB
a <u>M</u> aximum cost	<u>M</u> aximum pairwise distance	MMS	MMB

An optimal map classification that minimizes a total sum of costs can be formulated as follows:

$$\text{Minimize: } \sum_{i,j} c_{i,j} d_{i,j} \quad \forall i, j \in N, i < j \quad (3)$$

$$\text{Subjected to: } \sum_{i \in In_k} d_{i,k} = \sum_{j \in Out_k} d_{k,j} \quad \forall k \in N \quad (4)$$

$$\sum_j d_{s,j} = 1 \quad \forall j \in Out_s \quad (5)$$

$$\sum_i d_{i,e} = 1 \quad \forall i \in In_e \quad (6)$$

$$\sum_{i,j} d_{i,j} = T \quad \forall i, j \in N, i < j \quad (7)$$

$$d_{i,j} \in \{0,1\} \quad \forall i, j \in N, i < j \quad (8)$$

where $d_{i,j}$ is a binary decision variable, with one denoting that a line from i to j is part of a solution, and zero denotes otherwise; $c_{i,j}$ is the cost for a line from i to j ; In_k is a set of all nodes that start lines terminating at node k ; and, Out_k is a set of all nodes that terminate lines starting at node k . Nodes s and e respectively denote a starting node and an end node in a network; these nodes correspond to the smallest and largest values in the context of map classification. Equation (3) is the objective function for the model, which minimizes a total sum of costs. Equation (4) is a flow constraint to ensure that if a line terminates at a node, then the next line should begin from that node. Equation (5) ensures that the first node (i.e., the minimum value) participates in only the first line that begins from the first node (i.e., the first class). Similarly, equation (6) ensures that the last node (i.e., the maximum value) participates in only the last line (i.e., the last class). Equation (7) restricts the number of classes, and equation (8) is a binary integer restriction indicating whether or not each line is a part of a solution.

The second specification with the Minimax objective function can be defined as follows:

$$\text{Minimize:} \quad y \quad (9)$$

$$\text{Subjected to:} \quad \text{Equations (4) ~ (8)}$$

$$y \geq c_{i,j}d_{i,j} \quad \forall i, j \in N, i \leq j \quad (10)$$

This objective function is set with an auxiliary decision variable y , and equation (10) ensures that this decision variable achieves its maximum value.

3.3.4 An Optimal Map Classification Implementation

The proposed optimal classification models have been implemented as an extension of ArcGIS 10.1 (ESRI 2011) using C# in the Microsoft .Net Framework 4. The optimization problems are solved with Gurobi optimizer 6.5.0, which supports a branch-and-bound algorithm that is an exact method for finding an optimal solution. Figure 3-2 portrays the interface and graphic display of the map classification tool. In the right-hand side of the graphical user interface (GUI), all attribute estimates are displayed as blue dots in ascending order, together with their corresponding uncertainties that are represented as horizontal bars based on predefined confidence intervals. The vertical lines represent class breaks, and the colors of the bottom rectangles match the color symbols for classes of a choropleth map. This graphic display can be used to visually compare statistical differences of estimates, and evaluate a classification result. Observations in this graph dynamically link to features in a map so that a user can interactively explore data with their locations. Using the options in the left-hand side of the GUI, configurations of map classifications can be set; that is, a number of classes, optimal map classification strategies, and measure types (see Table 3-2 for configuration details). Also, a confidence level can be specified that determines the length of the horizontal bars in the graph. Finally, colors for a choropleth map can be controlled.

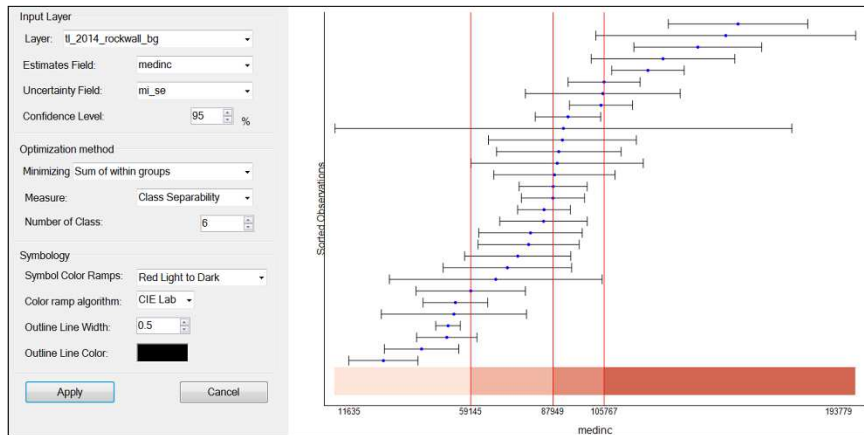


Figure 3-2. The GUI of the implemented tool

3.4 A Comparative Application

This section demonstrates the proposed optimal classification methods using the 5-year ACS data (2010-2014). This study directly uses ACS data, which contain MOE values that are calculated from the standard errors of the sampling distributions for survey estimates. A standard error (used to calculate a MOE) is a typical measure used to evaluate attribute uncertainty due to sampling error (Sun and Wong 2010). The statistical difference measures in this paper (i.e., the class separability measure and Bhattacharyya distance) can be calculated with standard errors as well as their estimates. Because ACS MOEs show the confidence bounds for a 90-percent confidence interval (U.S. Census Bureau 2009), the standard error of each estimate is derived directly from its MOE.

Specifically, the estimates of *median household income* (in 2014 inflation adjusted dollars) with their corresponding MOEs for the 254 counties in Texas are used in this application. Table 3-3 reports descriptive statistics for the dataset. The largest estimate of median household income is \$86,597, and the largest MOE is \$65,629. The largest MOE is observed for Loving County, with

a median household income of \$65,625. This MOE is extremely large compared with its estimate. The small population of Loving County (82 people and 39 households, according to the 2010 census) produces this extreme uncertainty. This county has a considerable influence on a map classification result, especially ones based upon the Bhattacharyya distance measure.

Table 3-3. Summary statistics of median household income (in \$) for the 254 counties in Texas

N	Estimate			MOE		
	Mean	Min	Max	Mean	Min	Max
254	46,352	22,176	86,597	4,576	338	69,245

Source: the 5-year (2010~2014) American Community Survey

Table 3-4 reports map classification results for the proposed methods; this table contains the range of observation counts³, and the total (i.e., class sum) and maximum (i.e., max) values. Underlined values are the lowest value in each column. Because an optimal classification method minimizes heterogeneity (or maximizes homogeneity) within a class, while incorporating uncertainty, the classification results of the proposed methods are compared with those of the class separability (Sun and Wong 2010) and the natural breaks classification methods. First, the class separability result has 247 for the range of observation counts, which indicates the classification result is highly unbalanced. That is, most of the observations are assigned to a single class. Figure 3-3 shows that the separability class breaks concentrate at lower values. This outcome is consistent with that for the separability method, which tends to produce a result in which class breaks tend to

³ The range of observation counts is calculated as the difference between the largest and smallest numbers of observations in classes.

be around extreme values (Sun, Wong, and Kronenfeld 2017). Thus, this classification method has difficulty achieving a balance in terms of the number of observations assigned to classes. The outcome based on a maximum pairwise distance for each class (i.e., SMS, SMB, MMS, and MMB; see Table 3-2 for acronym definitions) also shows relatively unbalanced results based on the range of observation counts in Table 3-4. In contrast, the optimal classification based on a sum of pairwise distances has an improved result in terms of balance. That is, SSS, MSS, SSB, and MSB in Table 3-4 have a more uniform distribution of numbers of observations across classes. The natural breaks classification achieves a relatively well balanced result in terms of the number of attribute estimates for each class (Table 3-4 and Figure 3-3), but considers only variances among attribute estimates while ignoring their uncertainties (Jenks 1977). The natural break classification result shows that counties forming large metropolitan areas in Texas (i.e., Dallas-Fort Worth, Houston, San Antonio, Austin) have high median household incomes. Some counties with high median household incomes also are located in the panhandle and west central regions of the state. In contrast, counties that are near the Mexican border tend to have low median household incomes. More counties with relatively low median household incomes are located in the eastern region of the state. The class separability map does not reveal this pattern, with most counties assigned to Class 6.

Table 3-4. A comparison of optimal classification methods with the class separability and the natural breaks classification methods

Methods	Range of observation counts	<u>Class separability measure</u>				<u>Bhattacharyya distance</u>			
		<u>Total</u>		<u>Maximum</u>		<u>Total</u>		<u>Maximum</u>	
		Class Sum	Class Max	Class Sum	Class Max	Class Sum	Class Max	Class Sum	Class Max
Natural Break	79	6235.04	58.09	3051.05	13.26	4192.1	155.22	1769.77	43.98
Class separability	247	109733.77	61.65	109733.67	61.55	258141.51	947.37	258141.36	947.22
SSS	38	<u>4544.7</u>	45.19	907.65	14.76	3586.15	111.06	842.85	54.51
SMS	175	31561.98	<u>32.31</u>	30915.77	17.33	32437.99	92.29	31857.88	75.22
MSS	34	4602.44	50.54	<u>788.28</u>	20.11	4176.32	157.7	1570.95	101.15
MMS	72	7095.79	48.72	2349.88	<u>10.54</u>	4927.31	108.94	1512.58	27.98
SSB	51	4811.36	49.34	1239.58	14.76	<u>3423.29</u>	127.97	713.91	54.51
SMB	131	13421.61	37.34	11482.10	10.54	9791.54	<u>73.95</u>	8075.92	27.98
MSB	47	4667.13	43.47	1184.55	14.76	3439.95	107.68	<u>604.19</u>	54.51
MMB	105	8635.96	49.92	5991.84	10.54	5859.46	109.78	3789.08	<u>27.97</u>

Note: (1) the best results are highlighted in bold and underlined.

(2) see Table 3-2 for the acronym definitions.

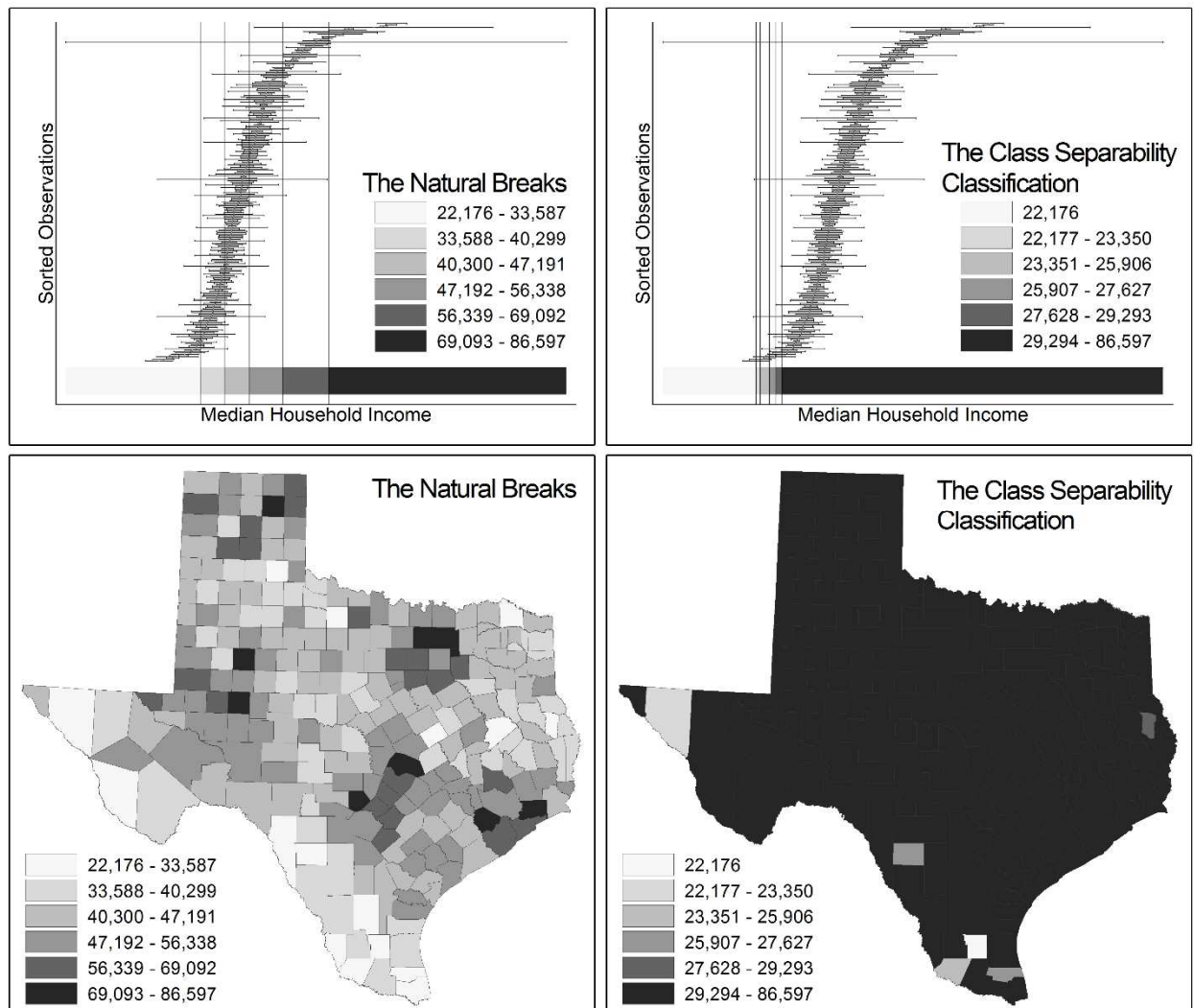


Figure 3-3. Classification results using the natural breaks and class separability methods

All optimal classification methods seek to produce homogeneous classes based upon different criteria (see Table 3-2), while often creating different spatial patterns of classification results. Figures 3-4 and 3-5 display the optimal classification results using the class separability measure. Among these, the SMS result is noticeably different from the others. The largest shift in groups occurs in the first group in the SMS result. In detail, 144 counties assigned to either Class 2, 3, or 4 in the SSS and MSS results are shifted into Class 1 in the SMS result. Only one county is assigned

to Class 3 in the SMS result, causing a large value for the range of observation counts. With a similar map classification result, the SSS and MSS criteria have a similar spatial pattern. Figure 3-6 portrays the changes of the map classification results from SSS to MSS, and then to SMS. Figure 3-6a displays the 15 counties that are classified differently according to the SSS and MSS criteria. In contrast, 214 counties in the SMS result map (Figure 3-6b) are classified differently than in the SSS result map, including the aforementioned 144 counties that are assigned to Class 1 according to the SMS criterion.

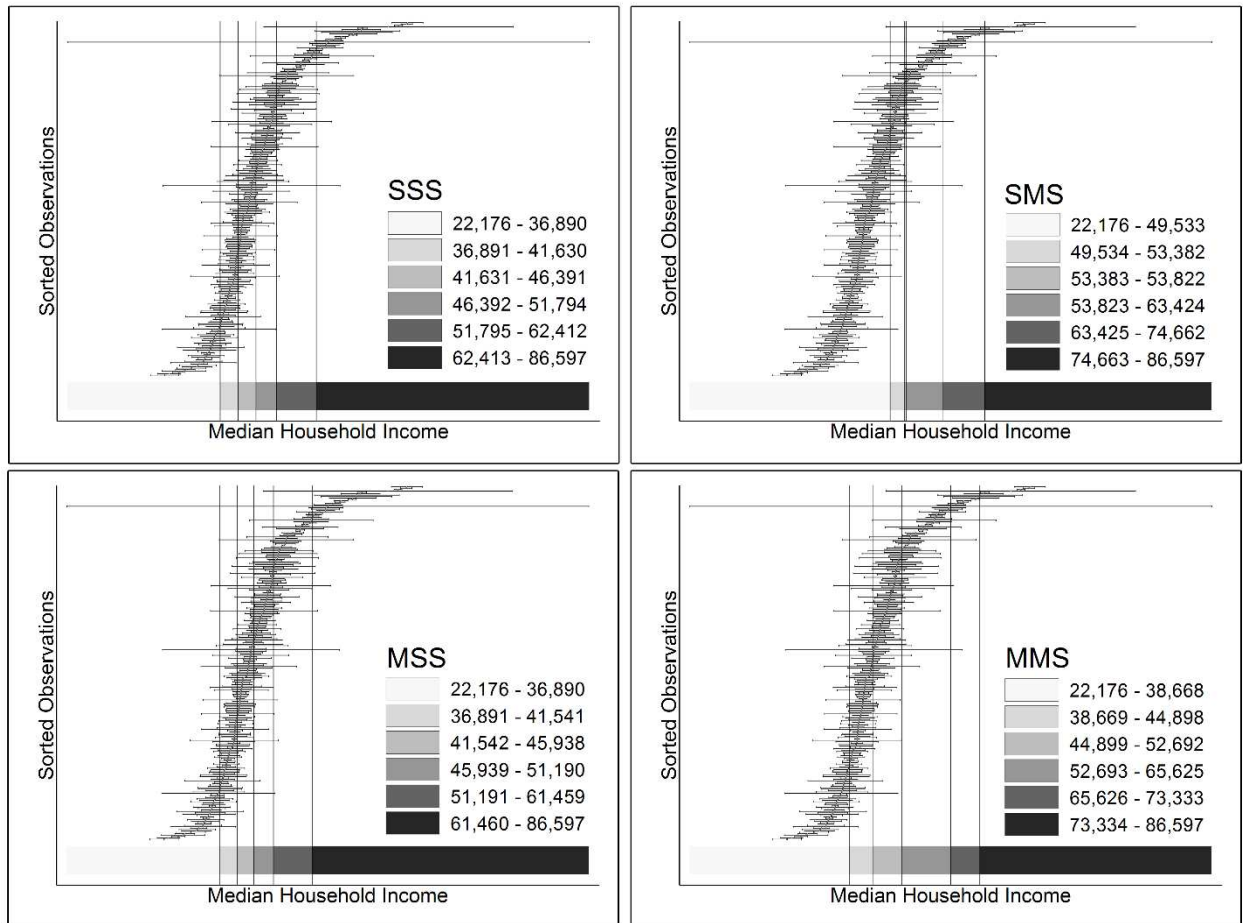


Figure 3-4. Optimal classification results by the class separability measure

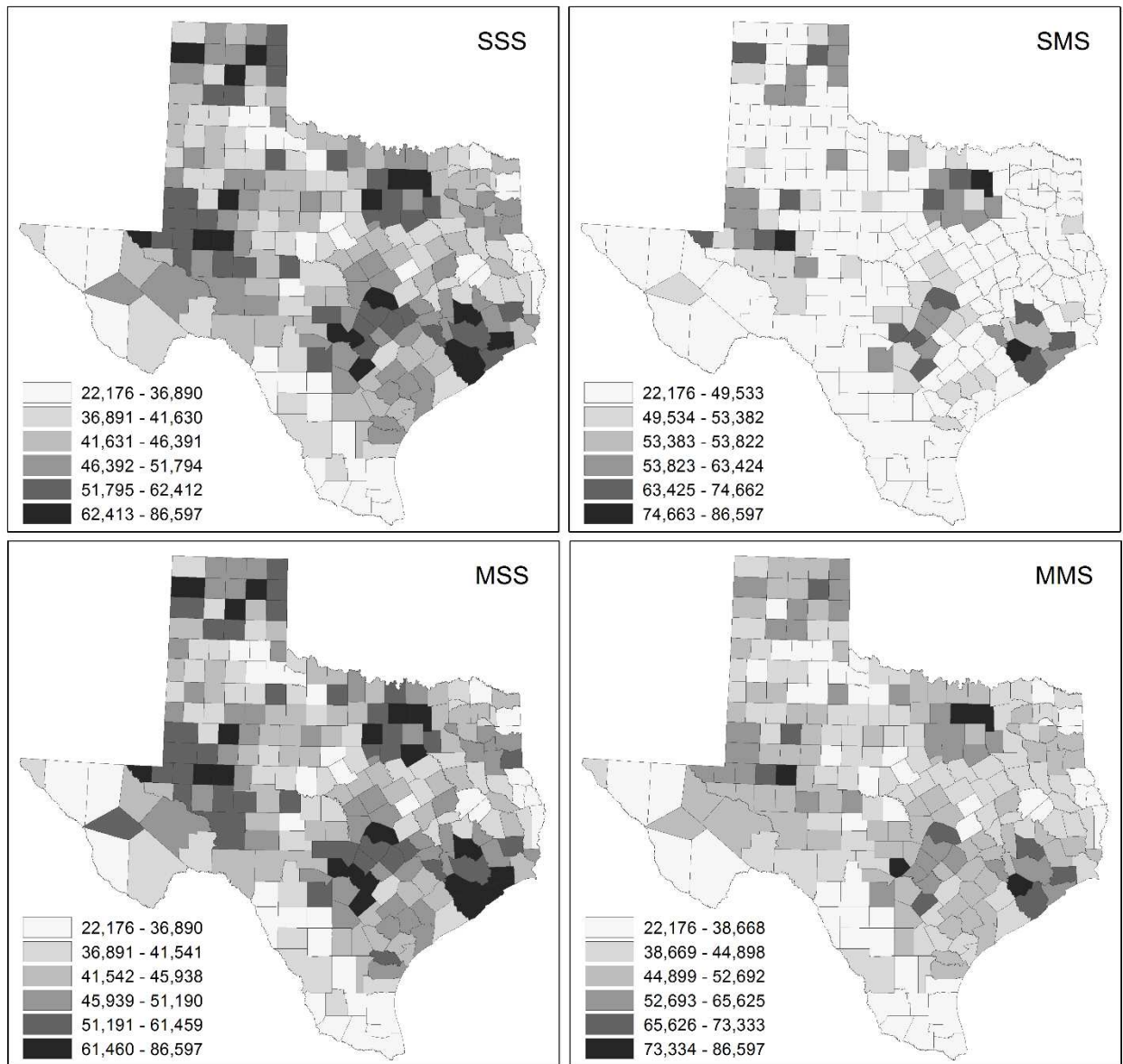


Figure 3-5. Optimal classification result maps using the class separability measure with different optimization criteria

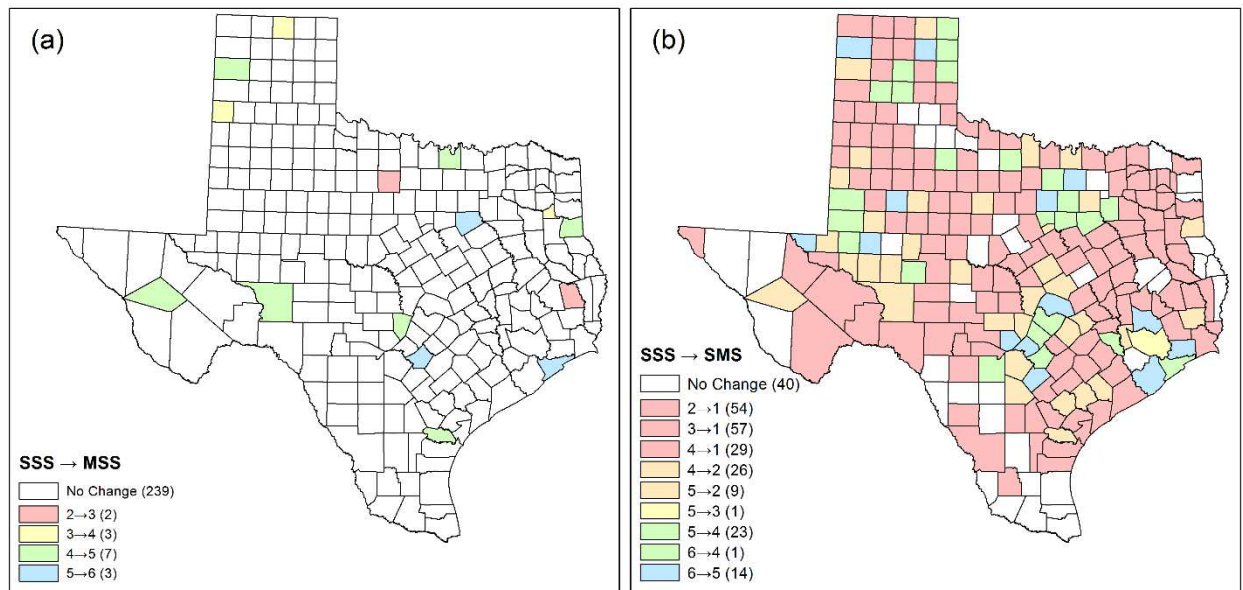


Figure 3-6. Class assignment differences using the separability measure: changes from SSS to (a) MSS, and (b) SMS

Figures 3-7 and 3-8 display the results of optimal map classification using Bhattacharyya distance. The SSB and MSB results are more similar than those of SMB and MMB. This outcome also may indicate that a choice of cost determination has more influence on a classification result than the objective functions. When Bhattacharyya distance results are compared with the class separability results, a noticeable difference can be observed with Loving County, which has the longest error bar in Figure 3-7. Because the Bhattacharyya distance measures a dissimilarity between probability distributions (Reyes-Aldasoro and Bhalerao 2006), the high standard error of Loving County can increase a distance measure. In contrast, its class separability measure, which is inversely proportional to its standard error, is smaller than Bhattacharyya distance, and has less impact on map classification. Figure 3-9 portrays differences in the map classification results between the SSB to MSB criteria. Figure 3-9a displays a similar result for SSB and MSB. Only 13

counties are classified differently in these two maps. In contrast, 154 counties are classified differently in the SSB and SMB maps (Figure 3-9b).

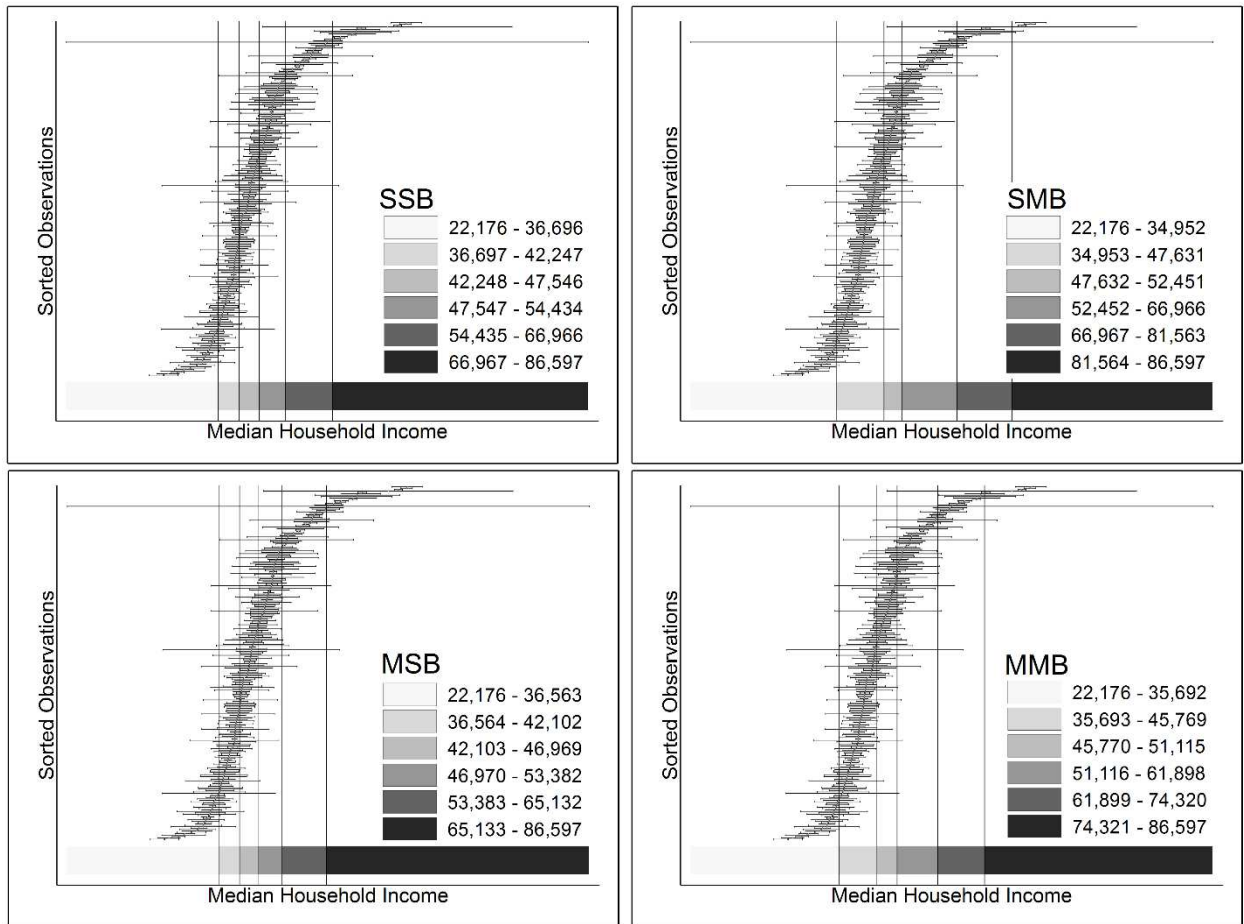


Figure 3-7. Optimal classification results based on Bhattacharyya distance

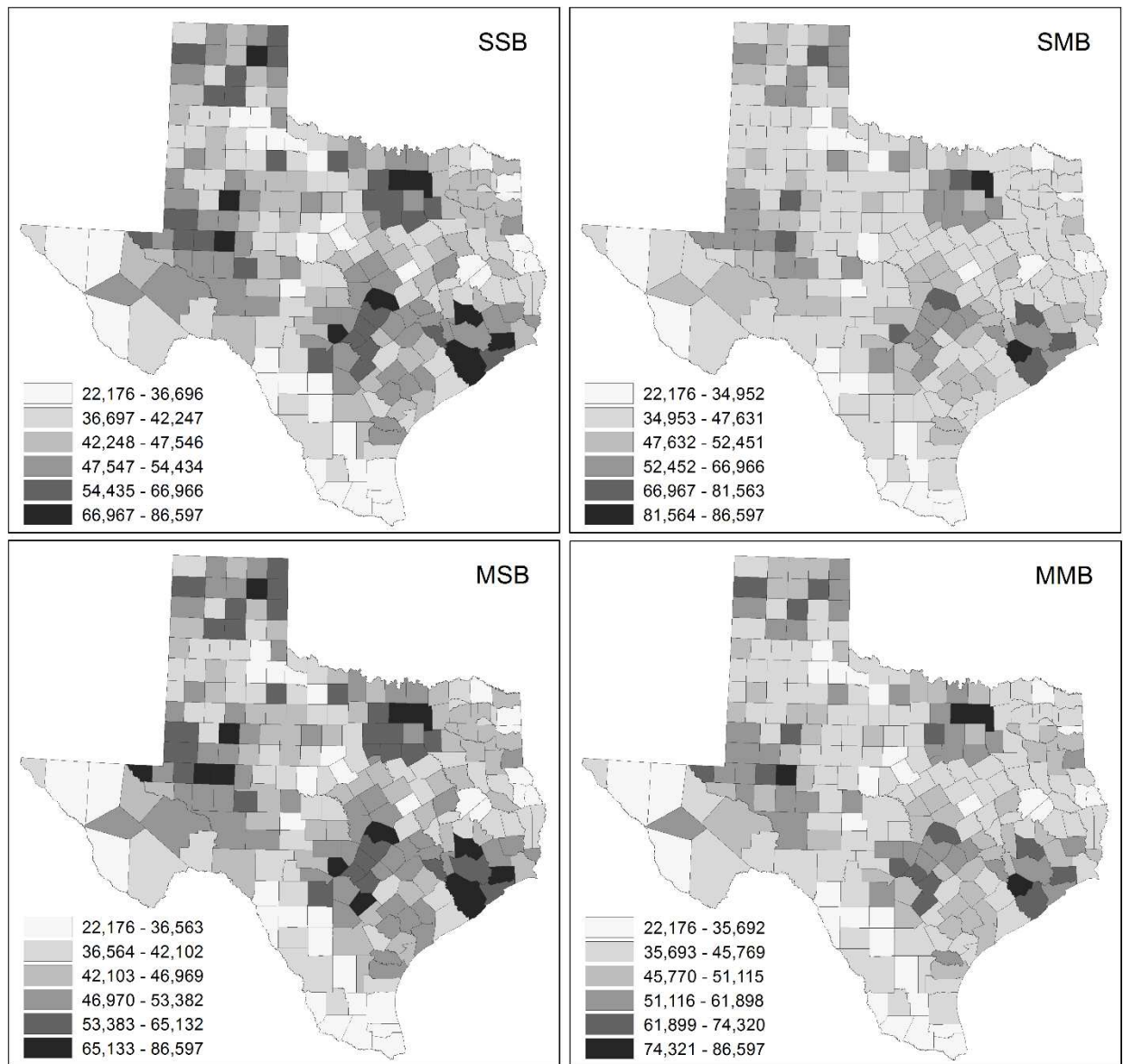


Figure 3-8. Optimal classification result maps using Bhattacharyya distance

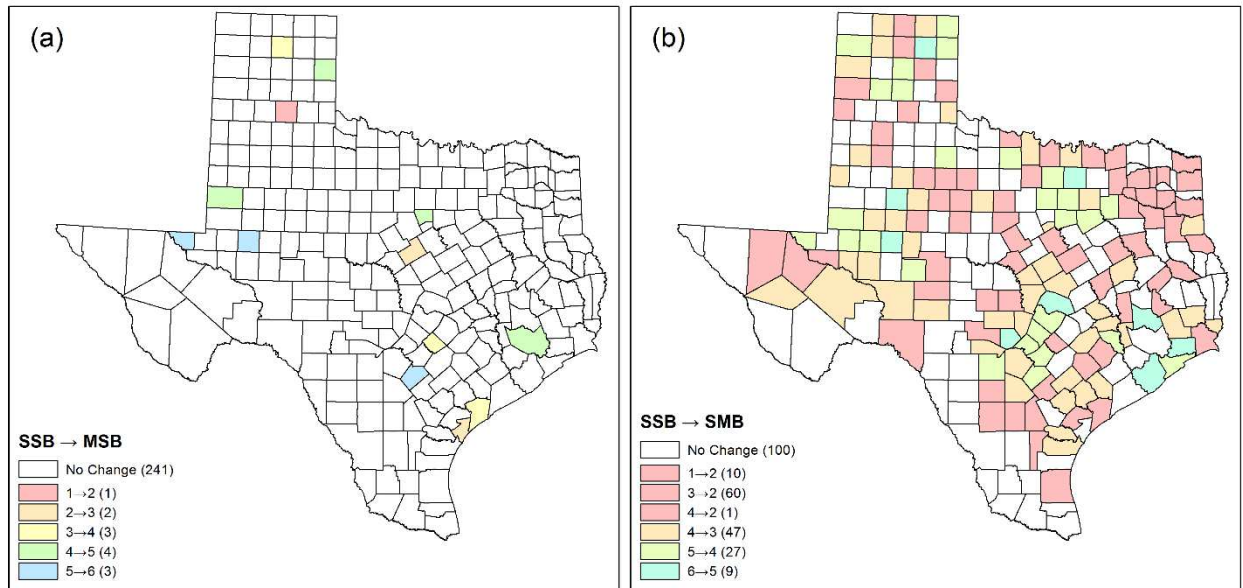


Figure 3-9. Class assignment differences using Bhattacharyya distance: changes from SSB to (a) MSB, and (b) SMB

3.5 Conclusions and limitations

This paper summarizes a study that investigated map classification by combining optimal classification methods (Cromley 1996) with various cost and objective functions that simultaneously consider attribute estimates and their uncertainties. Overall, the proposed classification methods successfully produce results achieving homogeneity within a class with a consideration of uncertainties as well as their attribute estimates. These methods also achieve a more balanced allocation of observations across classes than the class separability classification method. More specifically, the class separability measure and Bhattacharyya distance were used to compare statistical differences between a pair of observations. Also, costs were determined in two different ways, with objective functions minimizing either a total sum of costs or a maximum cost. The empirical application to 254 counties in Texas shows that cost determination has more

influence on the classification result than the objective function specification, and it can contribute to a more balanced number of observation among classes. In addition, a comparison of Bhattacharyya distance and separability results reveals that a noticeable difference between them can emerge for observations with greater uncertainty.

This paper contributes to map classification for thematic mapping in two ways. First, it extends map classification by incorporating uncertainty in an optimal classification. Previous approaches that incorporate uncertainty in map classification (Sun, Wong, and Kronenfeld 2014, 2017) often produce a less desirable result, with most observations assigned to a single class, and struggle to find an optimal result. Second, this paper presents a map classification tool that is implemented in a popular ArcGIS environment. This implementation furnishes a way to generate a thematic map that simultaneously considers attribute estimates and their uncertainties in its classification. This tool also can be used for an exploratory spatial data analysis that compares the underlying patterns of a spatial distribution while incorporating uncertainty information.

Some limitations to the proposed method merit further investigation in future studies. First, a performance evaluation (e.g., accuracy and robustness) for the devised optimal map classification with uncertainty needs to be investigated. Some common approaches are not necessarily effective in evaluating map classification while considering uncertainty. Jenks and Caspall (1971) suggest an overall accuracy index that combines individual indices that measure each objective of map classification. One weakness of this index is that the priorities of objectives still need to be assigned arbitrarily by map designers (Evans 1977). The robustness measure by Xiao, Calder, and Armstrong (2007) also has been used to evaluate the overall performance of classification methods (e.g., Traun and Loidl 2012; Sun, Wong, and Kronenfeld 2014). However, this measure only

evaluates the robustness of attribute estimates, and its tolerance level for the unreliability of a map classification is defined subjectively by a user. Second, the performance of the proposed map classification can be affected by the number of observations. A branch-and-bound algorithm solution can be extremely time consuming when the number of nodes in a branching tree is too large. Thus, as the number of observations increases, the time needed to find an optimal solution increases at an exponential rate. Third, the normality assumption for empirical data also can be problematic, although many non-normal variables can mimic, and hence are indistinguishable from, a bell-shaped curve in numerous circumstances (e.g., Poisson, binomial, beta, hypergeometric, gamma, chi-squared, Student t, F, negative binomial, and log-normal; see Griffith 2014). ACS attribute estimates are anticipated to be less sensitive to this normality assumption because the data provider with authority (i.e., the US Census Bureau) indicates that they follow a normal distribution (U.S. Census Bureau 2009). Designed based random sampling and the Central Limit Theorem furnish the justification for their contention. But the Central Limit Theorem is not always applicable for other empirical data, and fails to hold for a Cauchy random variable. A data transformation (i.e., Box-Cox power or Manly exponential transformation for normality; see Griffith 2013) may be necessary. Furthermore, although Bhattacharyya distance can be used for other distributions, the separability measure that utilizes the standard two-sample z-test may not be applicable for other distributions.

Acknowledgement

This research was supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this poster are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

CHAPTER 4
MODELING POSITIONAL UNCERTAINTY ACQUIRED THROUGH STREET
GEOCODING *

Authors - Hyeongmo Koo, Yongwan Chun, and Daniel A. Griffith

School of Economic, Political and Policy Sciences

The University of Texas at Dallas

800 West Campbell Road

Richardson, Texas 75080-3021

NOTE: This chapter is forthcoming in *International Journal of Applied Geospatial Research* in 2018: volume 9, issue 4.

Permission granted republish or display content in this dissertation.

ABSTRACT

Modeling positional uncertainty helps to understand potential factors of uncertainty, and to identify impacts of uncertainty on spatial analysis results. However, modeling geocoding positional uncertainty still is limited in providing a comprehensive explanation about these impacts, and requires further investigation of potential factors to enhance understanding of uncertainty. Furthermore, spatial autocorrelation among geocoded points has been barely considered in this type of modeling, although the presence of spatial autocorrelation is recognized in the literature. The purpose of this paper is to extend the discussion about modeling geocoding positional uncertainty by investigating potential factors with regression, whose model is appropriately specified to account for spatial autocorrelation. The analysis results for residential addresses in Volusia County, Florida reveal covariates that are significantly associated with uncertainty in geocoded points. In addition, these results confirm that spatial autocorrelation needs to be accounted for when modeling positional uncertainty.

Key Words: Positional uncertainty, geocoding, uncertainty modeling, spatial autocorrelation

4.1 Introduction

Uncertainty is inevitably embedded in any process involving measuring and representing the world, and uncertainty in spatial data and spatial data analysis has been extensively investigated in Geographical Information Sciences (Griffith, Wong, and Chun 2015). Positional uncertainty, which is one of five uncertainty types identified by the United States (U.S.) Federal Geographic Data Committee (FGDC), refers to inaccurate locations of geographic features (ANSI 1998). A common source of positional uncertainty is inaccuracy of the global positioning system (GPS) and geocoding errors. Horizontal and vertical positional accuracy of GPS points at reference stations have been reported regularly by the Federal Aviation Administration, suggesting that GPS recordings generally are reliable regardless of study area. In contrast, positional uncertainty of geocoded points can vary more, affected by geocoding algorithms and characteristics of study regions (Jacquez 2012), as well as properties of street networks (Zimmerman and Li 2010).

Understanding geocoding uncertainty is important to properly interpret spatial analysis results, which can be largely affected by the positional uncertainty of geocoded points (e.g., Burra et al. 2002; Harada and Shimada 2006; Griffith et al. 2007). Uncertainty can contribute to an increase in the standard errors of parameter estimates, and, subsequently, a reduction in the statistical power of a spatial cluster and/or trend detection (Zimmerman and Li 2010; Lee, Chun, and Griffith 2017). Previous studies (e.g., Cayo and Talbot 2003; Bichler and Balchak 2007; Zandbergen 2008a; Hart and Zandbergen 2013; Jones et al. 2014) emphasize that identifying geocoding errors can help researchers understand a geocoding error structure and set up an appropriate model specification for subsequent data analysis. These studies mainly focus on the magnitudes of geocoding uncertainty for various settings and locations. To achieve a reduction of uncertainty in a geocoding

process and spatial analysis results, uncertainty modeling often is considered as a prerequisite process (Zhang and Goodchild 2002), involving an exploration of possible covariates of the uncertainty. This investigation of possible covariates (i.e., properties of street networks) that potentially have an impact on the positional uncertainty of geocoded points can provide an insight into understanding geocoding positional errors. Those covariates can be utilized for identifying their impact on spatial analysis results (Zimmerman and Li 2010), and for predicting positional uncertainties at specific locations (Jacquez 2012). Interestingly, although Griffith et al. (2007) and Zimmerman, Li, and Fang (2010) show that the magnitude of geocoding positional uncertainty can be spatially autocorrelated, this data feature has received only limited attention in modeling positional errors of geocoded points (Zandbergen et al., 2012).

The purpose of this paper is to investigate and model positional uncertainty acquired through an automated geocoding process (which hereafter is called geocoding positional uncertainty). The main focus is on positional uncertainty in street geocoding (i.e., address matching), which is a predominantly automated geocoding process, at least in the United States (Zimmerman and Li 2010). First, this paper investigates the magnitudes of geocoding positional uncertainty. Second, this paper analyzes potential covariates for geocoding positional uncertainty and their impacts are examined in a regression context. Importantly, a proper theoretical probability distribution for geocoding positional uncertainty is investigated, and then the best distribution is utilized in a subsequent regression analysis. Third, this paper examines the structure of spatial autocorrelation in geocoding positional uncertainty, and properly accounts for spatial autocorrelation in modeling geocoding positional uncertainty. In this paper, 22,239 mailable residential addresses in Volusia County, Florida, 2016 are utilized for data analysis and modeling purposes.

4.2 Literature review

Geocoding positional uncertainty greatly influences results of a spatial analysis. Harada and Shimada (2006) examine the effect of geocoding positional uncertainty on kernel density estimation of crime locations, and conclude that the pattern of kernel density smoothing can be misrepresented by geocoding positional uncertainty. Similarly, Zinszer et al. (2010) show that kernel density estimation of disease distributions shows a discernible effect of geocoding positional uncertainty. DeLuca and Kanaroglou (2008) also demonstrate the impact of different geocoding methods on kernel density estimation, spatial scan statistics, and a bivariate K -function. Also, local spatial autocorrelation measures, including local Moran's I , G_i and G_i^* statistics, can be distorted by only a small amount of geocoding positional uncertainty (Burra et al. 2002). Griffith et al. (2007) evaluate the impacts of geocoding positional uncertainty on a spatial regression analysis of pediatric blood lead data, and report a clear impact on the spatial auto-binomial regression analysis results from geocoding positional uncertainty. Moreover, geocoding positional uncertainty has an influence even on spatial weights matrices because spatial weights are derived from the geocoded points (e.g., distance based weights and k -nearest neighbors) (Jacquez and Rommel 2009). Incorrect spatial weights lead to a substantially lower statistical power for spatial analyses (Anselin and Rey 1991).

Much previous research has investigated accuracy levels of geocoding positional uncertainty and its empirical distribution to understand its structural features. A comparative analysis for different places and settings has been popularly conducted in the literature. In detail, many studies compare geocoding positional uncertainty and the match rates of different geocoding methods that utilize different geocoding tools and/or algorithms (e.g., Karimi, Durcik, and Rasdorf 2004; Ward

et al. 2005; Zhan et al. 2006; Schootman et al. 2007; Strickland et al. 2007; Mazumdar et al. 2008; Jones et al. 2014; Chow, Dede-Bamfo, and Dahal 2015). In addition, geocoding positional uncertainty has been investigated in relation to address densities—i.e., urban and non-urban addresses (e.g., Bonner et al. 2003; Cayo and Talbot 2003; Ward et al. 2005; Mazumdar et al. 2008; Zimmerman and Li 2010)—and the impact of reference street network types (e.g., Whitsel et al. 2006; Zandbergen and Green 2007; Zandbergen 2008a, 2011). These comparative analyses provide insights into geocoding positional uncertainty, and help spatial scientists to understand and control uncertainty associated with geocoded points.

Exploring potential factors for geocoding positional uncertainty has drawn attention in uncertainty modeling. Generally, geocoding positional uncertainty is strongly related to the properties of corresponding street networks (Zimmerman and Li 2010) and geographical landscapes (Zandbergen et al., 2012) because the main components of geocoding positional uncertainty are errors in reference street networks and violations of a linear interpolation assumption (Jacquez and Rommel 2009). With regard to geographical landscapes, one well-known factor for geocoding positional uncertainty is whether or not an address lies in an urban or non-urban area (Zimmerman and Li 2010); places likely to have differences in their street segment properties. Geocoding positional uncertainty in non-urban areas generally is much greater than it tends to be in urban areas, because street segments in non-urban areas usually are longer and have fewer intersections (Cayo and Talbot 2003; Chow, Dede-Bamfo, and Dahal 2015). Geocoding positional uncertainty also gets larger when a small number of houses are not evenly distributed along a long street line, because a linear interpolation assumption, which means that house numbers are uniformly distributed along a street segment, does not hold (Levine and Kim 1998).

For this same reason, the length of a street segment and the density of street intersections have a significant influence on geocoding positional uncertainty (Zimmerman and Li 2010). However, these covariates provide only a partial explanation for positional uncertainty (Zimmerman and Li 2010), and more covariates need to be investigated for a better understanding of it in specific study areas (Jacquez 2012).

Positing an appropriate theoretical probability distribution is necessary to model geocoding positional uncertainty in regression. In early studies, error ellipses for point data based on a normal distribution was widely used to describe and visualize positional uncertainty (Dutton 1992), which also provides a basic model for positional uncertainty (Goodchild 1991; Wolf and Ghilani 1997). A comprehensive agreement about the single distribution describing geocoding positional uncertainty still does not exist (e.g., Zandbergen and Hart 2009), but a log-normal distribution is suggested to describe the empirical distribution of geocoding positional uncertainty in some studies, including Cayo and Talbot (2003) and Karimi et al. (2004). In contrast, Zandbergen (2008b) argues that geocoding positional uncertainty is not log-normally distributed. Zimmerman et al. (2007) claim that a mixture of bivariate t distributions is appropriate to describe geocoding positional uncertainty. A chi-square distribution also is suggested as a candidate theoretical distribution for geocoding positional uncertainty (Griffith et al. 2007). Although all of these previous studies have contributed to discovering a sampling distribution of geocoding positional uncertainty, their results still are limited to small geographical regions (Jacquez 2012).

Spatial autocorrelation commonly exists in geocoding positional uncertainty because of the aforementioned components of positional uncertainty (Zimmerman, Li, and Fang 2010). In other words, geocoded points that share the same reference information (e.g., their reference street

networks and geographical landscapes) tend to have a similar magnitude of positional uncertainty. Previous studies also report the presence of spatial autocorrelation in direction and magnitude of positional uncertainty in a digitizing process (Griffith 1989, 2008), and among individual geocoded points (Cayo and Talbot 2003; Zimmerman, Li, and Fang 2010), and in the numbers of misallocated geocoded points among census geographical units (Griffith et al. 2007). However, positional uncertainty frequently is assumed to be independent in its modeling (e.g., Zimmerman and Li 2010). That is, spatial autocorrelation components for geocoded points generally rarely have been accounted for in the literature.

4.3 Methods

This section describes the modeling procedure of geocoding positional uncertainty. First, it presents the process of input data preparation and geocoding positional uncertainty quantification for two different types of reference points. Then, it discusses theoretical probability distributions that have been recognized in the literature for modeling geocoding positional uncertainty. Finally, it presents potential covariates and model specifications for both non-spatial and spatial regression to examine impacts of these covariates on geocoding positional uncertainty.

4.3.1 Data

This paper uses 22,236 mailable residential addresses, which are a subset of all 238,706 addresses in Volusia County, Florida, in 2016. These street addresses are obtained in the form of an address point database, in which an address point for every occupied building in this county is

stored as a point location, typically placed at a building's center (Zandbergen, 2011). The input addresses are classified as urban or non-urban using a standard point-in-polygon operation based on the 2015 Urban Areas polygons created by the U.S. Census Bureau. To control for any effects of different types of addresses (e.g., residential and commercial addresses) on positional uncertainty, only residential addresses are utilized here. Residential addresses are identified through a standard point-in-polygon operation with an ancillary zoning dataset available from Volusia County⁴. Figure 4-1 shows the study area and urban boundaries as well as residential addresses. A street reference dataset is obtained from Volusia County that has street centerlines at a scale of 1:4,800; this dataset is constantly updated and considered to be the most accurate source of street reference data for this county.

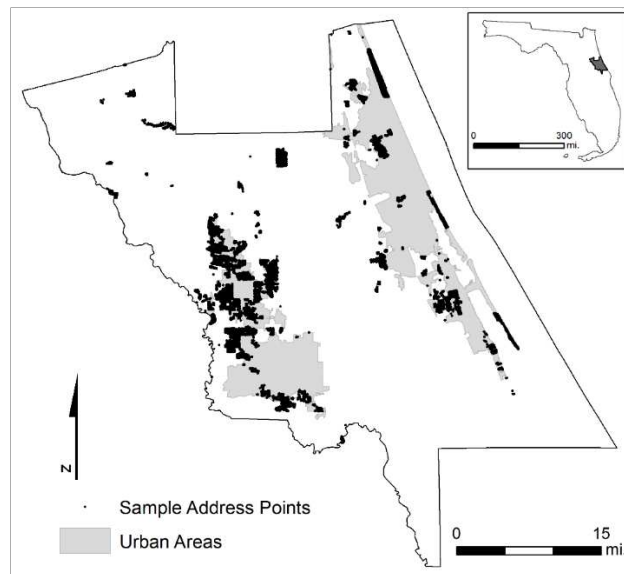


Figure 4-1. The selected addresses in Volusia County, Florida in 2016

⁴ <http://maps.vcgov.org/gis/download/shapes.htm>

4.3.2 A geocoding process and positional uncertainty quantification

Geocoded points are obtained by perfect matching addresses to the local street centerlines from Volusia County using the automated geocoding tool in ArcGIS 10.1 (ESRI 2001). This tool begins by separating address lists into address components, such as house number and street name, and then standardizes these components (Zimmerman et al. 2007). Next, this process compares these components with address-ranged street segments based on a match score that measures the similarity between addresses and candidate segments. This match score is normalized to numbers between 0 and 100. Only perfectly matched geocoded points (i.e., the match score: 100) are used in order to reduce uncertainty arising from incomplete addresses. Next, the coordinates of the geocoded points along the street segments are obtained by linear interpolation between starting and ending points, which are defined by the address ranges of corresponding street segments. No offset value is given for both sides and ends from street centerlines to eliminate the effect of offset values.

Geocoding positional uncertainty is quantified based on two different types of reference points, which are building centers and parcel centroids. First, the building centers were obtained from the address point dataset in which address points typically are located at the centers of buildings (Zandbergen, 2011). Furthermore, all locations of the input addresses were thoroughly examined using Google Earth, and the locations of the input addresses were moved to the centers of their corresponding buildings based upon Google Earth. Second, parcel centroids were obtained from parcel data from Volusia County. These input addresses were relinked to their corresponding parcels by the parcel identification numbers (PIDs) in the parcel dataset. Each parcel centroid is

calculated using a center of the gravity based algorithm (Okabe and Sugihara 2012, p61). Figure 4-2 shows an example of these two types of reference points.

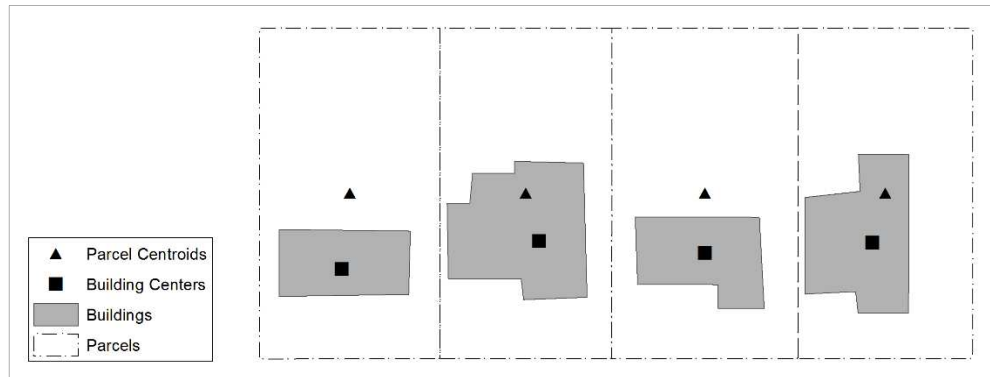


Figure 4-2. Reference points to measure positional uncertainty

4.3.3 Distribution of geocoding positional uncertainty

Identification of a proper probability distribution model for an empirical dataset often is necessary for a modeling exercise (Delignette-muller and Dutang 2015). Although a proper distribution for geocoding positional uncertainty still needs further investigation (Zandbergen and Hart 2009), most study findings agree that the empirical distribution does not always follow a normal distribution (Zimmerman et al. 2007). In this paper, several probability distribution models are investigated. First, a log-normal distribution is considered because it has been utilized in previous research (e.g., Karimi, Durcik, and Rasdorf 2004; Zandbergen 2008b). Second, an exponential distribution, which is a special case of a gamma distribution (Stacy 1962), is considered as another candidate, because this probability model is widely used to represent geographic distance decay effects (e.g., Griffith and Wong 2007). Thus, this study examines these two candidate theoretical probability distribution models as well as a gamma distribution (i.e., the

generalization of the exponential distribution). Parameters for a candidate distribution are estimated with maximum likelihood (MLE). This distribution fitting procedure is conducted separately for the two different reference point sets for the urban/non-urban classification. The robustness of this fitting result is examined with bootstrapping.

4.3.4 Regression model specifications

Several covariates that can potentially be associated with geocoding positional uncertainty are examined with regression and a proper probability distribution model. Both non-spatial and spatial regression models are considered, and their results are compared. In spatial regression, geographic neighbors are specified with the k -nearest neighbors (KNN) criterion. In order to choose a suitable k for the KNN weights, an experiment was conducted with a series of k values from 1 to 20. A k value is determined based on the level of spatial autocorrelation detected and model performance indices. Specifically, a k value was chosen that had the greatest residual spatial autocorrelation, the smallest AIC, and/or the greatest pseudo- R^2 value⁵. In addition, spatial autoregressive (SAR) models were employed to account for spatial autocorrelation in geocoding positional uncertainty.

Covariates can be classified into two broad categories, which are the properties of street networks and geocoding locations (Table 4-1). The properties of street networks that are reference data in a geocoding process have an influence on geocoding positional uncertainty (Zimmerman and Li 2010). Studies (Cayo and Talbot 2003; Zimmerman and Li 2010; Chow, Dede-Bamfo, and

⁵ Pseudo- R^2 values are calculated with the formula given by Nagelkerke (1991)

Dahal 2015) found that positional uncertainty is relatively large when street segments are long and have few intersections. Thus, the lengths of street segments are included, and are measured directly from the reference street networks. The densities of street networks at geocoding locations are considered because these densities can reflect the urbanization status of geocoding location. The literature reports that geocoding positional uncertainty is smaller in urban than non-urban areas (Cayo and Talbot 2003; Ward et al. 2005; Mazumdar et al. 2008). That is, geocoding positional uncertainty is expected to decrease along with an increase in street network density. Street network density and the smoothness of estimated density values are markedly affected by the search radius employed in density estimation (Silverman 1986). In this paper, the search radius is set to 3,000 feet, a quantity based upon the average size of parcels in the study area. Also, the ratio between the count of housing units and the number of allocated addresses is included because a violation of the linear interpolation assumption might be exacerbated when there is a great difference between these two quantities. If the count of housing units is much less than the number of allocated addresses, geocoding positional uncertainty is expected to increase. In a geocoding process, addresses with even and odd-numbers fall on opposite (e.g., the left and right) sides of a street segment, and the location of a geocoded point is linearly interpolated based on each side of a street segment. Thus, the number of allocated addresses and the count of housing units also are prepared separately for each side of a street segment, which makes the covariate values differ depending on the side of an address, even if they are on the same street segment.

Table 4-1. Covariates for positional uncertainty in geocoding, together with their abbreviations

Street networks	Length of street segments (<i>length</i>), street network density (<i>density</i>), and the ratio of the count of housing units and on the number of allocated addresses (<i>house</i>)
Geocoding locations	Places containing geocoding locations (<i>urban</i>), relative geocoding locations vis-à-vis the corresponding street segments (<i>location</i>), and geocoding location parcel sizes (<i>parcel</i>)

In addition to street network properties, the characteristics of geocoding locations on street segments are considered as covariates. First, a relative geocoding location on its matching street segment is considered. That is, this designated location represents how close a geocoding location is to the midpoint of a street segment. A relative location to a segment midpoint (r) is calculated as:

$$r = \begin{cases} p \times 2 & p \leq 0.5 \\ (1 - p) \times 2 & p > 0.5 \end{cases}, \text{ where } p = \frac{t-h}{t-f}.$$

Here, p denotes the ratio of, in the numerator, an address number of the geocoding location (h) subtracted from the last address number (t) of the corresponding street segment to, in the denominator, a range of allocated addresses that is the last address number minus the first address number (f). If p is greater than 0.5, p is subtracted from 1. Then, r is obtained by multiplying p by 2 in order for r to range from 0 to 1. A small r value indicates that a geocoding location is close to either the start or end point of a street segment, whereas a large value indicates a geocoding location close to the midpoint of the street segment. This relative location is calculated for each side of a street segment because the range of addresses can differ by street side. Second, a dummy variable for urban (0) and non-urban (1) is included. As discussed earlier, geocoding positional uncertainty in urban areas tends to be smaller than in non-urban areas (Zimmerman and Li 2010).

Statistical interaction terms between this dummy variable and the other covariates also were explored. Finally, the parcel size of geocoding locations might have an association with geocoding positional uncertainty. These parcel sizes were obtained from the Volusia County parcel dataset, and then linked to matching addresses using their PIDs. Because residential buildings with a large size parcel generally are far from a street centerline, a large parcel size might be positively associated with geocoding positional uncertainty.

4.4 Results

This section provides the modeling results of geocoding positional uncertainty. It reports the descriptive statistics of geocoding uncertainty for the urban and non-urban parts of the study area, and then the fitting results of the empirical frequency distributions of geocoding positional uncertainty. Finally, non-spatial and spatial regression results are presented in order to examine the effects of potential covariates (Table 4-1).

4.4.1 Descriptive statistics

Geocoding positional uncertainty is quantified using Euclidean distance from either of the two reference point types: building centers, and parcel centroids. Table 4-2 presents the descriptive statistics of positional uncertainty for the sample urban and non-urban areas. The total number of input residential addresses is 22,236, with 19,757 being urban, and 2,479 being non-urban addresses.

Table 4-2. Descriptive statistics for positional uncertainty of sample geocoded points

Categories	Reference points	Sample Size	Minimum	Median	Maximum	Mean	Standard deviation
Non-urban	Building	2,479	12.86	56.81	665.00	74.28	58.33
Urban	Building	19,757	6.30	34.60	350.10	42.45	22.93
All	Building	22,236	6.30	36.02	665.00	46.00	30.77
Non-urban	Parcel	2,479	14.71	65.24	664.70	81.60	59.97
Urban	Parcel	19,757	12.75	37.54	488.70	45.44	23.38
All	Parcel	22,236	12.75	39.12	664.70	49.47	31.88

The positional uncertainty measured using building centers ranges from 6.30 to 655.00 meters, with a median of 36.02 meters, and a mean of 46.00 meters. The substantial difference between urban and non-urban addresses coincides with findings in the literature (e.g., Cayo and Talbot 2003; Whitsel et al. 2006). The mean and median for the non-urban area are, respectively, 74.28 and 56.81 meters, which are much larger than their urban area counterparts of 42.45 and 34.60 meters, respectively. Overall, positional uncertainty measured when using parcel centroids is slightly larger than when using building centers, which ranges from 12.75 to 664.70 meters, with, respectively, a mean and median of 49.47 and 39.12 meters.

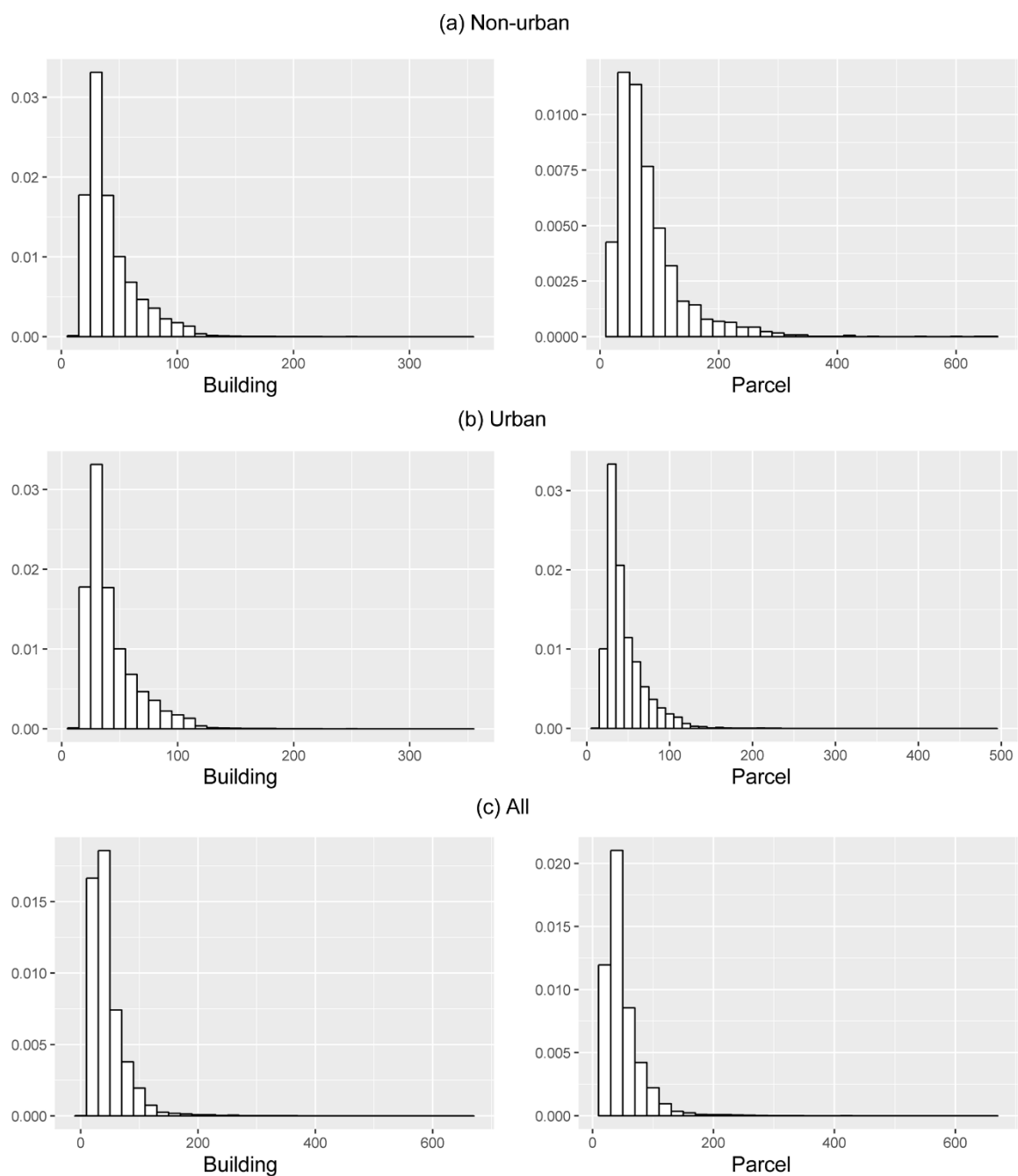


Figure 4-3. Histograms of the geocoding positional uncertainties for the different reference points

Figure 4-3 displays histograms of positional uncertainty for the urban and non-urban areas. All histograms show a strong positive skewness; that is, they are skewed to the right side. These positional uncertainties do not show considerable differences in their distributions between the two sets of reference points; i.e., building centers, and parcel centroids, for the same type of geographic landscape. Note that zero offset values are used for both sides from street centerlines, which means that all geocoded points are located on street centerlines. Therefore, the positional uncertainties from building centers and parcel centroids to the geocoded points should contain a small amount of uncertainty caused by these offset values. This data feature can explain a number of observations close to zero; i.e., the first bin, in these histograms. This data feature is likely to make these histograms conform more closely to a log-normal than a gamma distribution, as found in previous studies (e.g., Cayo and Talbot 2003; Karimi, Durcik, and Rasdorf 2004; Zandbergen 2008b).

4.4.2 Distribution fitting results

The empirical distributions of geocoding positional uncertainty are fitted with three candidate probability distribution models: the log-normal, exponential, and gamma distributions. Figure 4-4 presents the empirical cumulative distributions of geocoding positional uncertainty superimposed on their fitted theoretical cumulative distribution counterparts. This figure shows that empirical geocoding positional uncertainty is closer to a log-normal distribution than to the other two, for both the urban and non-urban areas under study.

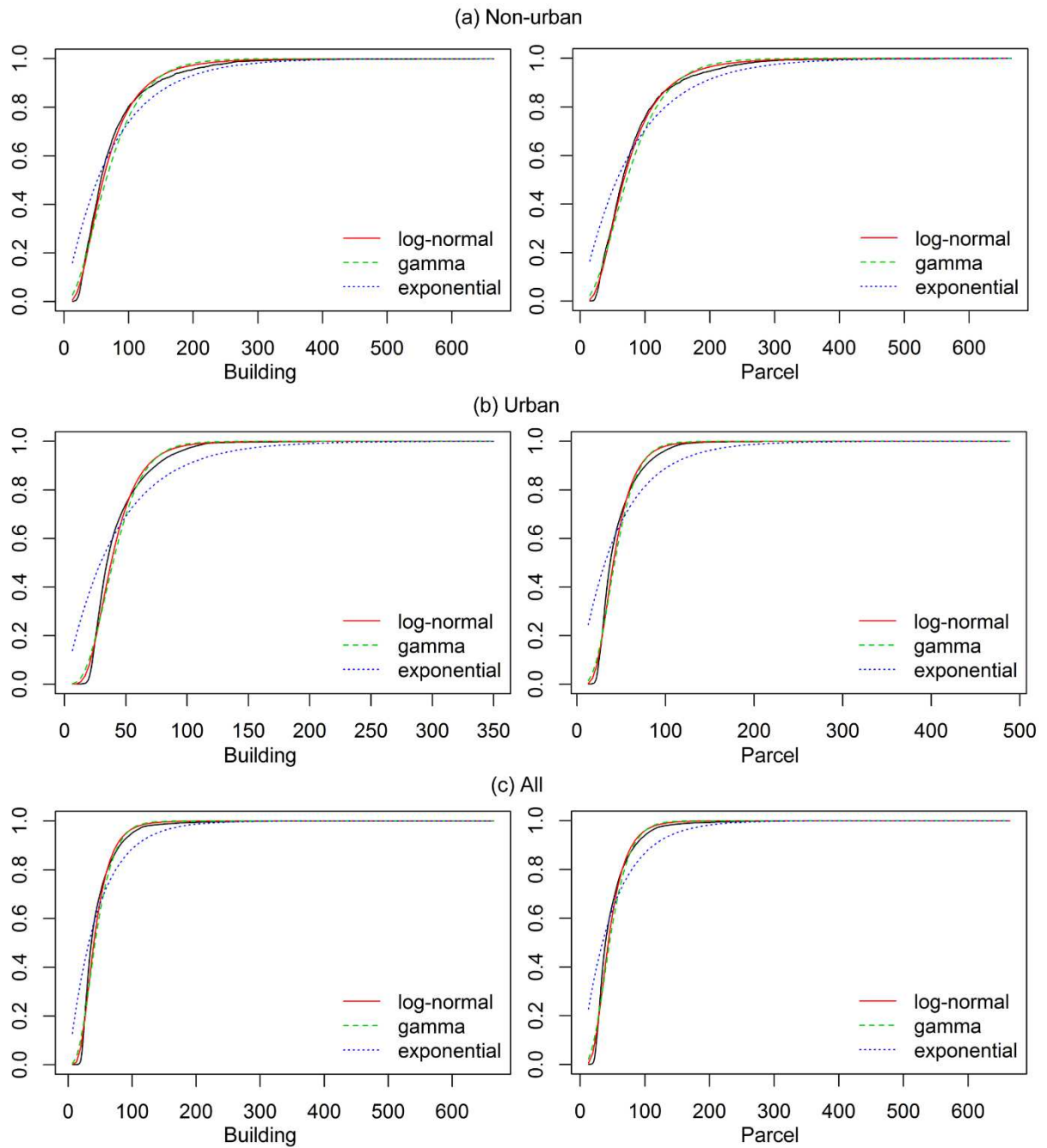


Figure 4-4. The empirical cumulative distributions of geocoding positional uncertainty superimposed on their theoretical distribution counterparts (i.e., log-normal, exponential, and gamma distributions)

The closeness of these distribution pairs is further examined with a Kolmogorov-Smirnov (K-S) statistic, and a Bayesian information criteria (BIC). The K-S statistic measures the distance between an empirical and a reference cumulative distribution functions (Stephens 1986), whereas the BIC consists of measures of badness-of-fit and model complexity (Zimmerman et al. 2007). Thus, a distribution with a smaller K-S statistic and a smaller BIC is preferable. Table 3 reports the calculated K-S and BIC results. According to these statistics, positional uncertainty for both sets of reference points most closely conforms to a log-normal distribution. This finding corroborates that from a visual inspection of the histograms in Figure 4-3. Substantial differences in the fitting results are not found between urban and non-urban addresses for the same reference type (i.e., building, parcel). This result may suggest that separate regression models are not necessary for the urban/non-urban areas.

Table 4-3. The Kolmogorov-Smirnov (K-S) and the Bayesian information criteria (BIC) diagnostic statistics

Categories	Reference points	K-S statistic			BIC		
		log-normal	gamma	exponential	log-normal	gamma	exponential
Non-urban	Building	0.045	0.092	0.237	30,872	31,277	32,215
Urban	Building	0.083	0.110	0.350	215,212	217,809	234,586
All	Building	0.084	0.114	0.328	248,597	252,754	267,590
Non-urban	Parcel	0.029	0.070	0.238	31,279	31,614	32,681
Urban	Parcel	0.083	0.108	0.358	216,523	218,999	237,271
All	Parcel	0.086	0.112	0.335	250,835	254,895	270,821

The robustness of this preceding distribution fitting is examined with bootstrapping. Based on the K-S statistics (Figure 4-5), the results of the simulation experiment are consistent with the

previous result. That is, a log-normal distribution furnishes the best description of geocoding positional uncertainty for both reference types. Also, no noticeable difference is identified between urban and non-urban areas for the same reference type.

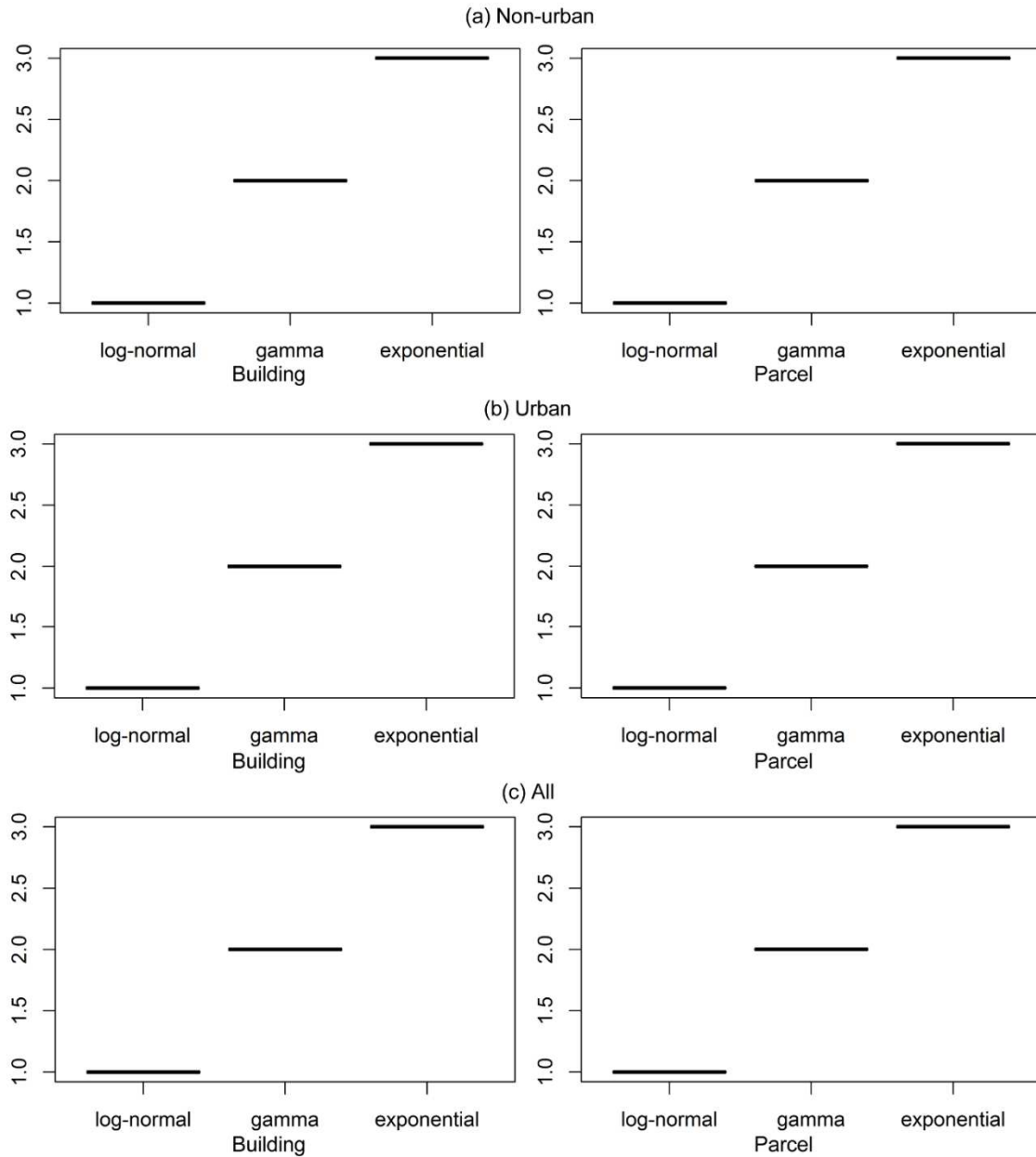


Figure 4-5. Boxplots of the Kolmogorov-Smirnov (K-S) statistic values from bootstrapping

4.4.3 Regression results

The effects of the covariates (Table 4-1) on geocoding positional uncertainty are investigated using linear regression models because a log-normal distribution is preferred based upon the preceding results. For linear regression models, dependent variables are defined as the log-transformed positional uncertainty measured from building centers (LM-Building) and parcel centroids (LM-Parcel). Spatial autocorrelation in the regression residuals is measured with the Moran coefficient (MC), in which spatial neighbors of geocoded points are determined by the KNN with $k = 2$.⁶

Table 4-4 reports the results of the linear regression analysis. For LM-Building, all covariates have a significant association with positional uncertainty measured from building centers. The positively significant variables at the 0.1% level are *length* (0.758), *urban* (0.310), and *parcel* (0.175). In contrast, *density* (-0.005), *location* (-0.233), and *house* (-0.247) have a significant negative relationship at the same level. These three variables have different associations for the urban/non-urban classification based on the estimates of the interaction terms with *urban*: that is, the t values of *density:urban*, *location:urban*, and *parcel:urban* are -8.177, 12.052, and -10.027, respectively. LM-Parcel has the same significance results as LM-Building at the 0.1% level (Table 4-4). One noticeable difference is the increase in the *parcel* coefficient estimate (0.175 for LM-Building, increasing to 0.239 for LM-Parcel). The reason for this increase is rather straightforward. The geocoded points in this analysis are located on street center lines due to zero offset values, but parcel centroids generally are farther from street centerlines as a parcel size increases. LM-Parcel

⁶ The KNN weight with $k = 2$ yields the largest MC value for values of k between 2 and 20.

has a better model fit with a higher adjusted R^2 value (0.348) than LM-Building (0.263). Coefficient estimates for the statistical interaction terms with the urban/non-urban dummy variable also are consistent with those for LM-Building. The confidence intervals for the coefficient estimations derived from bootstrapping with both linear regression models coincide with observed coefficients of the covariates, in terms of their direction and magnitudes (See Appendix B1 and B2).

Table 4-4. Linear regression model results: LM-Building, and LM-Parcel

Covariates	LM-Building		LM-Parcel	
	Coefficient	t-value	Coefficient	t-value
Intercept	3.689	284.932***	3.719	314.948***
Length	0.758	31.983***	0.690	31.896***
Urban	0.310	7.896***	0.328	9.152***
Density	-0.005	-7.195***	-0.003	-4.696***
Location	-0.233	-23.378***	-0.189	-20.775***
House	-0.247	-25.735***	-0.281	-32.140***
Parcel	0.175	33.352***	0.239	50.130***
Length:Urban	-0.066	-1.370	0.005	0.114
Density:Urban	-0.033	-8.177***	-0.035	-9.494***
Location:Urban	0.360	12.051***	0.298	10.925***
House:Urban	-0.062	-1.453	0.045	1.147
Parcel:Urban	-0.074	-10.026***	-0.097	-14.407***
Adj. R^2	0.263		0.348	
F-stats.	723		1,078	
(p-value)	(< 0.001)		(< 0.001)	
MC of	0.663		0.676	
Residuals	(< 0.001)		(< 0.001)	
(p-value)				

Note: Significance codes: ***0.001, **0.01, *0.05

Variance inflation factor (VIF) values for all variables without interaction terms are less than 10.

However, these models only explain modest proportions of variances in their corresponding dependent variables, with adjusted R^2 values of 0.263 for LM-Building and 0.348 for LM-Parcel. Although most covariates are significantly related to geocoding positional uncertainty, these levels of significances may be boosted by the large number of observation (i.e., 22,236) input addresses (Lin, Lucas, and Shmueli 2013). In addition, the residuals of both models show a significant level of positive spatial autocorrelation; the residual MC values are 0.663 and 0.676, respectively (the p -values are less than 0.001). These models suffer from spatial autocorrelation and need to be extended to appropriately account for it.

4.4.4 Spatial regression model results

Table 4-5 summarizes the results of the SAR models with a spatial weights matrix determined by the KNN weight ($k = 2$). Again, the SAR result with $k = 2$ has the highest pseudo- R^2 value and the lowest AIC among k values from 2 to 20. Like the previous linear regression models, two SAR models are specified with the two different dependent variables: the log-transformed positional uncertainty measured from building centers (SAR-Building), and from parcel centroids (SAR-Parcel). The SAR analysis furnishes some noticeable improvements to those obtained with the linear regression models. First, the estimates of the spatial autocorrelation parameter are significant. Values greater than 0.6 indicate a high level of positive spatial autocorrelation. That is, positive spatial autocorrelation that persists in the linear regression residuals is successfully accounted for in these SAR specifications. Second, both SAR models show better model fits than the linear regression models. Specifically, the SAR models have much higher pseudo- R^2 values

(0.582 for SAR-Building, and 0.641 for SAR-Parcel). That is, this result indicates that the variance of the positional uncertainty is better explained in the SAR specifications.

Both SAR models yield the same results as the linear regression models in terms of statistical significance for the covariates. That is, *length*, *urban*, and *parcel* are positively, and *density*, *location*, and *house* are negatively, associated with the dependent variables. However, the estimated coefficients for each variable are slightly different. For the SAR-Building specification, *length* and *parcel* show a significantly positive relationship (0.698 and 0.101, respectively), which indicate greater geocoding positional uncertainty tends to coincide with longer street segments and larger parcels of address points. Also, a dramatic increase in the *parcel* coefficient estimate (to 0.169) is found in the SAR-Parcel results, when compared to its linear regressions counterpart. In addition, *house* has a significant association with its dependent variable (z-value=-21.589) in the SAR-Parcel. This outcome suggests that geocoding positional uncertainty tends to be larger when the count of housing units is much less than the number of allocated addresses on a street segment. The negative associations for *location* (-0.166 for SAR-Building, and -0.128 for SAR-Parcel) also are significant in both SAR models. The estimated coefficients for *location* indicate that larger positional uncertainties tend to occur more at starting or ending points of a street segment than at its midpoint. *Urban* and *density* are also significantly associated with geocoding positional uncertainty. The same result is reported in previous studies (e.g., Cayo and Talbot 2003; Zimmerman and Li 2010), which indicates that urban areas with a higher density of street networks have smaller positional uncertainty than non-urban areas with a lower density of street networks. Appendix B3 and B4 illustrate the confidence intervals for the coefficient estimations from bootstrapping for both SAR models, indicating robustness of the estimated coefficients.

Table 4-5. The results of SAR models: SAR-Building and SAR-Parcel

Covariates	SAR-Building		SAR-Parcel	
	Coefficient	z-value	Coefficient	z-value
Intercept	3.763	193.758***	3.767	212.448***
Length	0.698	21.265***	0.624	20.939***
Urban	0.289	5.545***	0.310	6.551***
Density	-0.009	-6.846***	-0.006	-4.890***
Location	-0.166	-18.783***	-0.128	-16.162***
House	-0.268	-19.454***	-0.270	-21.589***
Parcel	0.101	23.707***	0.169	44.470***
Length:Urban	-0.085	-1.410	0.014	0.255
Density:Urban	-0.022	-4.203***	-0.020	-4.283***
Location:Urban	0.225	8.977***	0.188	8.402***
House:Urban	-0.026	-0.415	-0.027	-0.477
Parcel:Urban	-0.020	-3.272***	-0.055	-9.925***
Pseudo-R ²	0.582		0.641	
Lambda	0.629		0.641	
(p-value)	(< 0.001)		(< 0.001)	

Note: Significance codes: ***0.001, **0.01, *0.05

Variance inflation factor (VIF) values for all variables without interaction terms are less than 10.

Among the statistical interaction terms with *urban*, in both models, the three significant variables (*density:urban*, *location:urban*, and *parcel:urban*) have different effects on geocoding positional uncertainty in urban and non-urban areas. While *density:urban* (-0.022 for SAR-Building, and -0.020 for SAR-Parcel) has the same relationship nature as *density* (-0.009 for SAR-Building, and -0.006 for SAR-Parcel), *location:urban* (0.225 for SAR-Building, and 0.188 for SAR-Parcel) and *parcel:urban* (-0.020 for SAR-Building, and -0.055 for SAR-Parcel) have the opposite relationship nature. That is, the impact of *density* is stronger in non-urban than urban areas, whereas the influence levels of *parcel* and *location* are weaker in the non-urban area.

Moreover, *location* has an opposite association in the urban and non-urban areas. In the urban area, as mentioned earlier, *location* is negatively associated with geocoding positional uncertainty, whereas this variable shows a positive association in the non-urban area. This result shows that positional uncertainty tends to be larger around the midpoint of a street segment, rather than near the starting or ending points of a segment in a non-urban area.

4.5 Conclusions

Positional uncertainty acquired through an automated geocoding process (what is referred to as geocoding positional uncertainty in this paper) is examined based upon residential addresses in Volusia County, Florida. This paper summarizes findings for measuring geocoding positional uncertainty from two types of reference points, namely building centers and parcel centroids. Although overall geocoding positional uncertainty measured from parcel centroids is slightly larger than from building centers, substantial differences are not found between the different reference types in the distribution fitting results. Using regression techniques, including linear and SAR specifications, an extensive set of covariates that relate to the properties of street networks and geocoding locations on street segments were examined. Linear regression and SAR model results show that all of the selected covariates are significantly associated with geocoding positional uncertainty. Although both linear and SAR model results have the same signs for the estimated coefficients, the SAR model has improved model performance because it accounts for spatial autocorrelation in geocoding positional uncertainty. Coefficient estimates for the statistical interaction terms suggest that impacts of three variables differ between urban and non-urban areas.

These three variables are street network density (*density*), relative geocoding locations on the corresponding street segments (*location*), and parcel sizes of geocoding locations (*parcel*).

This paper contributes to the literature in two ways. First, this paper identifies multiple covariates that rarely have been investigated, including *location* and *house*, and examine them with the well-known factors of geocoding positional uncertainty, namely, *length*, *parcel*, and *density*. The regression results show that the newly suggested factors have a significant association with geocoding positional uncertainty, in addition to the factors that have been examined in the literature. Thus, this paper contributes to improving geocoding positional uncertainty modeling by incorporating the additional factors. Also, this paper confirms that effects of some covariates (i.e., *density*, *location*, and *parcel*) on geocoding positional uncertainty are significantly different between urban and non-urban areas, which provides further explanation of the relationship between geographical landscapes housing geocoding positional uncertainty, and its related covariates. Second, this study extends positional uncertainty modeling by considering spatial autocorrelation. The presence of spatial autocorrelation in positional uncertainty has been recognized (e.g., Griffith 1989, 2008; Cayo and Talbot 2003; Griffith et al. 2007; Zimmerman, Li, and Fang 2010), but it has not been considered in modeling geocoding positional uncertainty. This paper empirically confirms the presence of spatial autocorrelation in geocoding positional uncertainty. It further shows that spatial autocorrelation should be considered when conceptualizing and analyzing geocoding positional uncertainty.

This research can be extended in the future. First, the input addresses obtained from the address point database are well structured and thoroughly cleaned while such well-prepared data are rarely available for most empirical analyses. The input addresses are highly standardized and well

maintained, which results in higher match rates (Zandbergen, 2011). Furthermore, because this paper uses only perfectly matched locations to mitigate this effect, the positional uncertainty might be much smaller than for other empirical datasets. Second, the analysis in this paper is conducted using a specific, small geographic region. Hence, its findings may not be readily applicable to other geographical regions. The covariates suggested in this paper might have different associations with geocoding positional uncertainty in other study areas. Nevertheless, the covariates are expected to provide a good starting point for the more general modeling of positional uncertainty in a geocoding process.

Acknowledgements

This research is supported by the National Institutes of Health, grant 1R01HD076020-01A1; any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Institutes of Health.

CHAPTER 5

CONCLUDING REMARKS

This dissertation contributes to the literature of spatial data uncertainty in two broad ways. First, it provides implementations for geovisualization and map classification methods incorporating spatial data uncertainty. Although various aspects of spatial data uncertainty have been widely investigated (e.g., Shi, Fisher, and Goodchild 2002; Zhang and Goodchild 2002), dealing with spatial data uncertainty is very limited in a general GIS environment due to a lack of implementations. In this respect, this dissertation can enhance the usability of spatial data uncertainty for general GIS users. Specifically, the implementation of attribute uncertainty geovisualization can help GIS users to utilize many visual variables that have been investigated for their uncertainty geovisualization in the literature. In addition, the developed tool for an optimal map classification can assist to incorporate uncertainty information in a thematic mapping process, and furthermore, a map classification result can be evaluated and compared with other classification results by the graphic diagnostic tool (Figure 3-2).

Second, this dissertation extends the literature of spatial data uncertainty by providing methods to manage and reduce spatial data uncertainty. The literature furnishes a comprehensive discussion about the sources of spatial data uncertainty (e.g., ANSI 1998), the way of its quantification (e.g., Bonner et al. 2003; Zimmerman and Li 2010), and impacts of spatial data uncertainty on analysis results (e.g., Griffith et al. 2007; Lee, Chun, and Griffith 2017). But, investigations about methods to control spatial data uncertainty essentially are lacking. In this regard, the modeling of positional uncertainty from a geocoding process allows anticipating geocoding positional uncertainty based on the properties of street networks and geocoding locations, which might help to reduce and

predict the impact of geocoding positional uncertainty on spatial analysis results. In addition, the optimal map classification developed in this dissertation can increase reliability of a thematic map pattern by considering attribute uncertainty in a mapping process. Therefore, the modeling of geocoding positional uncertainty and optimal map classification with attribute uncertainty contribute to the literature by providing an example of controlling spatial data uncertainty.

The followings provide additional explanations, and discusses limitations and further research for the individual sub-topics. The first sub-topic for attribute uncertainty visualization is necessary to address two limitations. First, although this dissertation mainly uses probability theory for uncertainty representation (e.g., coefficient variances and standard errors), various options, including Bayesian techniques and the Dempster-Shafer theory, also can be utilized. Uncertainty visualization with these representations is worth addressing. Especially, a fuzzy set approach is highly applicable to represent uncertainty because of its ability to display partial memberships. However, all types of data cannot be represented using this approach. For example, bell curves around a statistical estimate (e.g., the margins of error in the American Community Survey data) are not good candidates for this fuzzy set theory (Davis and Keller 1997). Second, a focused group survey with various samples and target groups may offer a more objective evaluation of the effectiveness of the implemented visualization methods. Some studies in the literature (e.g., Goodchild, Battenfield, and Wood 1994; Slocum et al. 2009) discuss the visual variables utilized in this implementation, and the effectiveness of bivariate mapping for uncertainty visualization. Specifically, MacEachren et al. (2012) show that color value leads to a better intuitive outcome to represent uncertainty than color saturation in their empirical study. However, still a group survey

would provide useful insights for evaluating the effectiveness of the developed tools with various geographical scales and target users.

The second research topic deals with map classification in the presence of uncertainty information. Currently, various mapping tools support unclassified maps, which have been promoted because of their accurate reflection of a real-world distribution. However, advances in unclassified maps does not make a traditional classified map and map classification useless. It is still widely used because of its own merits and the demerits of unclassified maps. The first and the most critical disadvantage of unclassified maps is a limited ability of human eyes to differentiate colors (or symbols) in a continuous scale (Slocum et al. 2009). Colors (i.e., shades) on unclassified maps are accurate proportionally to the value of each spatial unit, although this mathematical accuracy might not coincide with perceptual accuracy. That is, matching colors on an unclassified map with its legend is not easy due to enormous numbers of different colors in the map and the simultaneous contrasts on a color from its neighbors. In addition, when a skewed distribution of data exists, the ordinal relationship in the data might be concealed. Thus, a classified map and a map classification method are still important in cartography. These reasons contribute to the development of map classification incorporating uncertainty information.

In addition to an unclassified map, a way to determine an appropriate number of classes in the proposed optimal map classification methods might be worth investigating. Like other optimal map classification methods (e.g., Jenks's natural breaks method), the optimal methods proposed in this dissertation also have objective functions and values that represent homogeneity within classes incorporating uncertainty information. These objective values can be used to assist in selecting an appropriate number of classes. Generally, an objective value is expected to decrease

as the number of classes increases. Thus, an appropriate number of classes can be chosen by exploring the contribution of class counts to the objective values using a plot and/or sensitivity analysis.

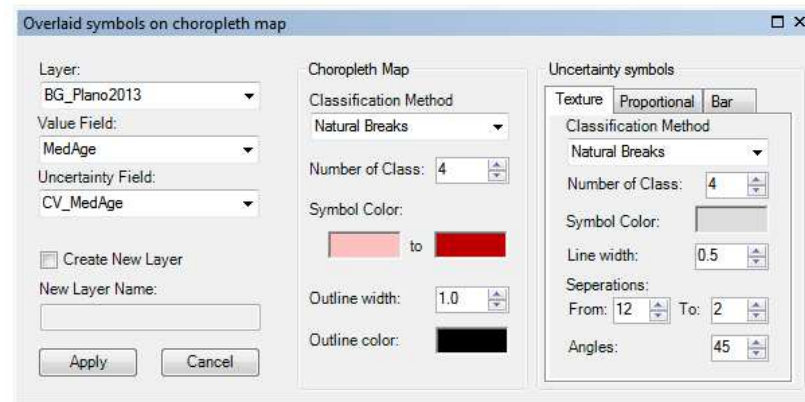
The third research topic shows geocoding uncertainty modeling with an empirical dataset: 22,236 mailable residential addresses in Volusia County, Florida. All locations of these input addresses were thoroughly examined and refined using Google Earth satellite images. Specifically, 3,368 address points obtained from an address point database do not have associated with actual buildings on the satellite images. All of these addresses were excluded in this analysis because building centers that are used as reference points cannot be quantified. Note that most of the address points are correctly located at the centers of their corresponding buildings on Google Earth, although high resolution Google Earth images also have horizontal positional uncertainty (e.g., Pulighe, Baiocchi, and Lupia 2016). However, 1,837 address points show a huge discrepancy from their building centers, and these points were moved to their corresponding building centers based on Google Earth images. Furthermore, parcel data, from which one type of reference points (i.e., parcel centroids) and one covariate are derived, were also refined. Specifically, 164 address points are located in multi-part parcels, which generally are divided by streets, might yield inaccurate parcel centroids and sizes. Thus, individual multi-part parcels also were examined with Google Earth images, and then converted to single-part polygons so that appropriate parcel centroids and their areas are generated.

APPENDIX A

THE GRAPHIC USER INTERFACE

A1. GUI for overlaid symbols on a choropleth map (OSCM)

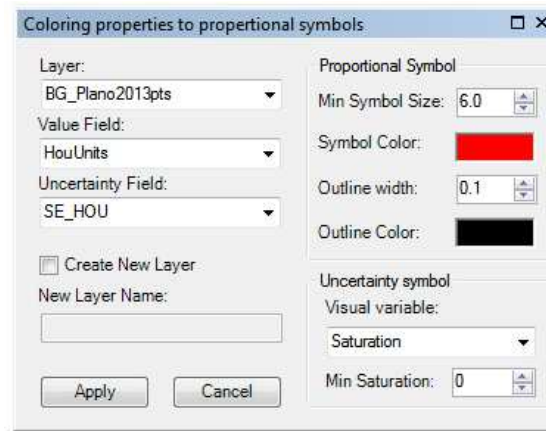
The graphical user interface (GUI) has three major sections. The first section is to specify a target layer, an attribute field, and a corresponding uncertainty field in the target layer. This part is identical in the GUIs for CPPS and CS. The second section is to configure an output choropleth map: that is, the number of classes and symbols. Note that the color ramp is automatically created with a start and an end color. In the third section, a user can choose a type of overlaid symbol with corresponding tab pages, each of which has its own options for uncertainty symbolization.



A2. GUI for coloring properties to proportional symbols (CPPS)

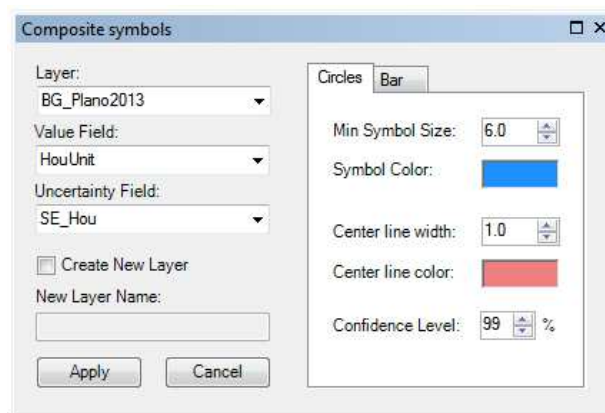
The first section is same as the one in the GUI for OSCM. The second section is to specify attributes and their uncertainty visualization. Symbol sizes are determined with mathematical scaling based on a minimum symbol size, which a user should provide. Uncertainty is symbolized with two different visual variables, color saturation and color value in the hue-saturation-value

(HSV) color model, which are listed in a combo box. The saturation or color value is linearly calculated from that of a selected color to a chosen minimum level.



A3. GUI of composite symbols (CS)

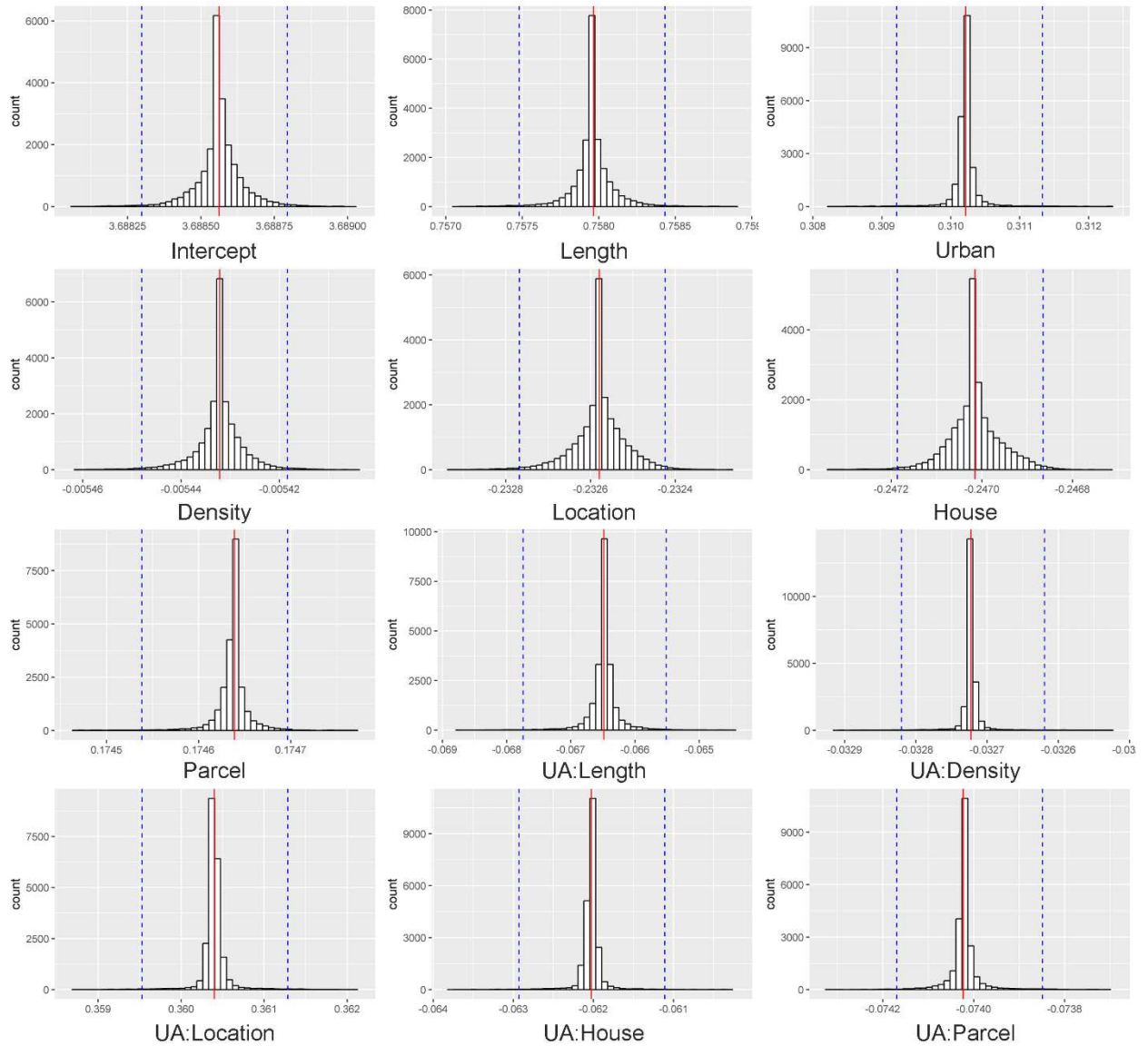
The first section is the same as the one for the other GUIs. An uncertainty field needs to contain standard errors. With the second section, different composite symbols can be chosen with tab pages; i.e., a circle or a bar. A confidence level should be specified to determine the symbols size of the confidence interval.



APPENDIX B

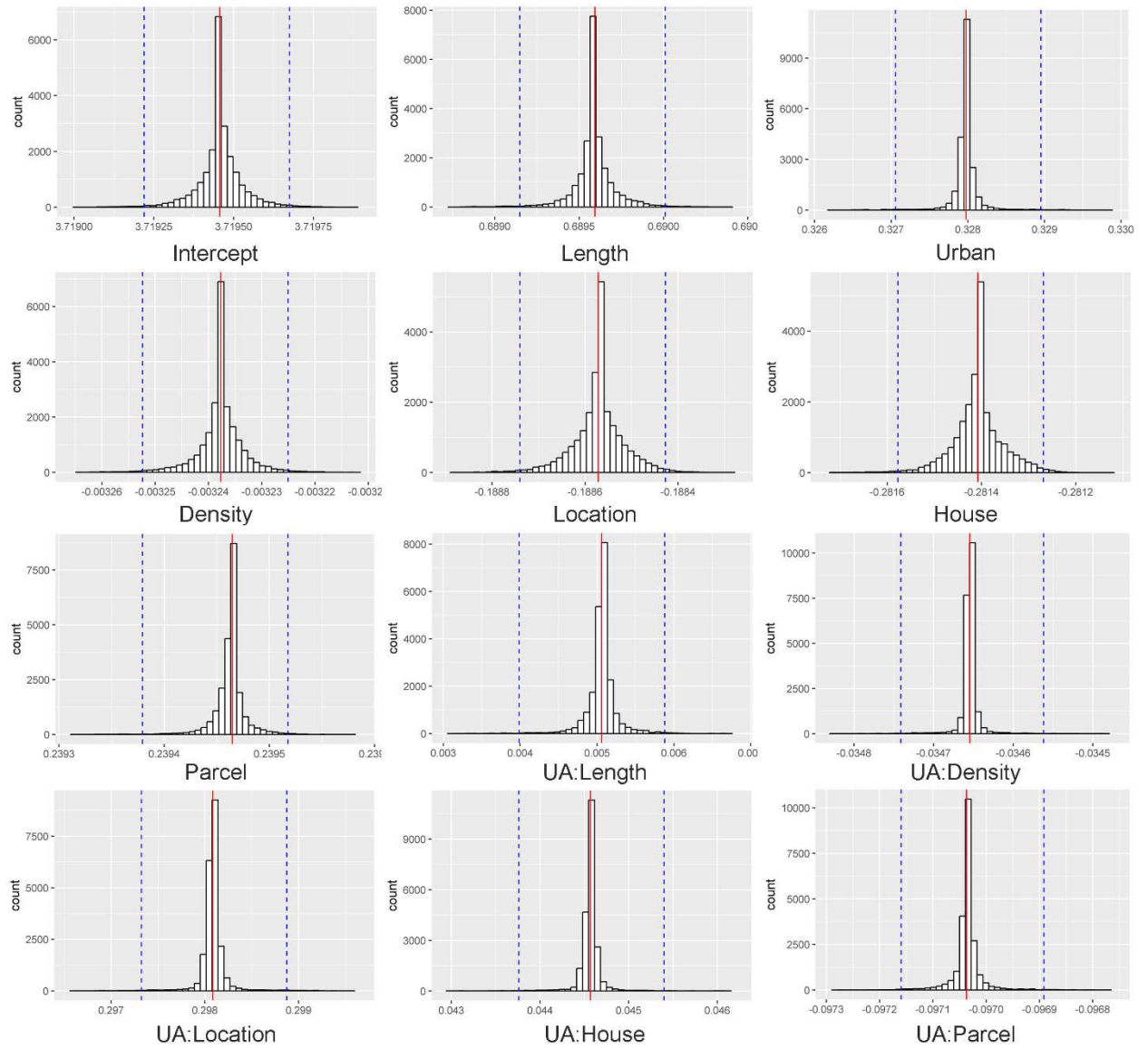
BOOTSTRAPPING RESULTS

B1. Bootstrapping results of LM-Building



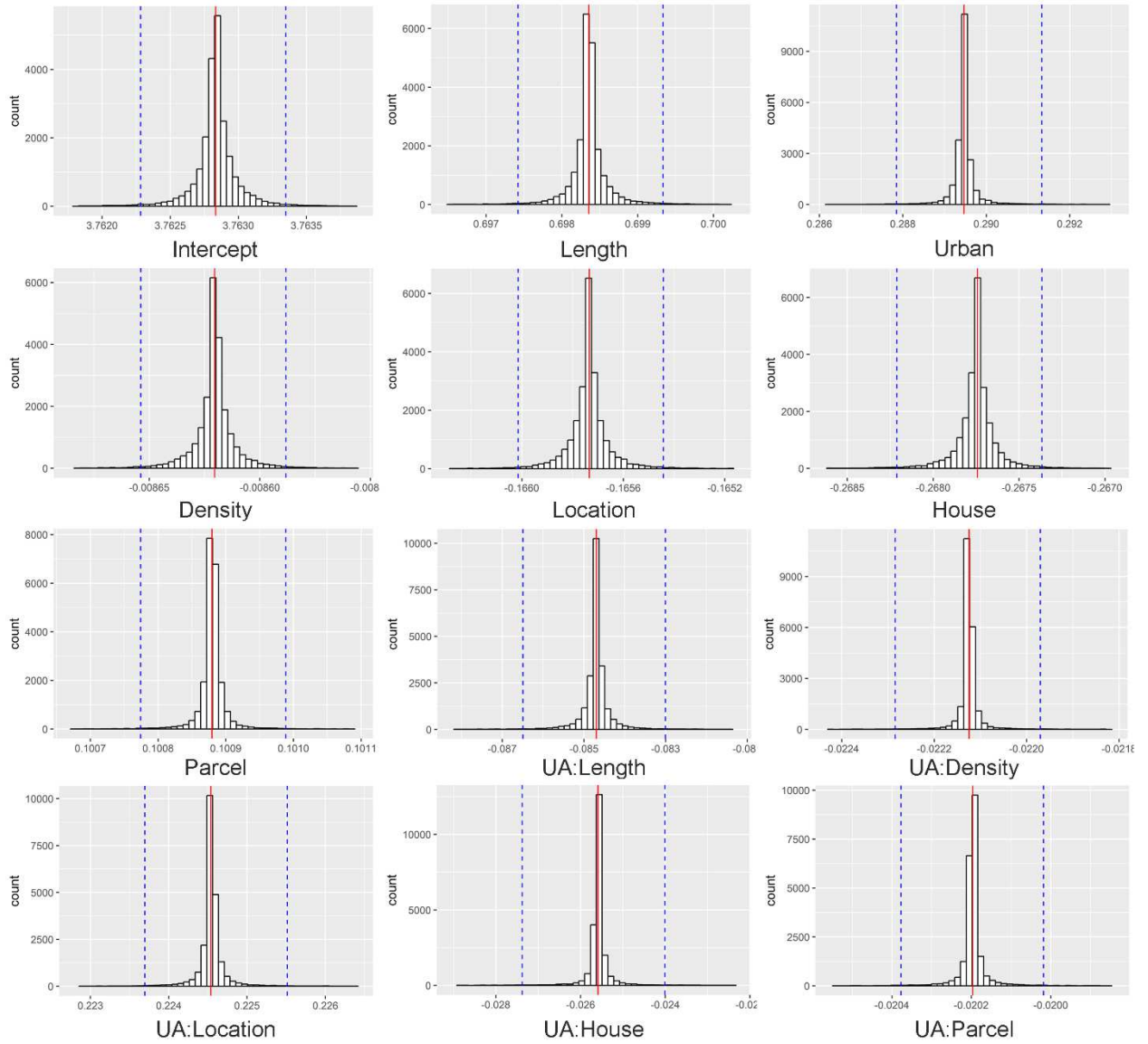
Note: the dashed vertical lines display the confidence intervals of the corresponding coefficients for a 99 percent confidence level, and the red vertical lines mark the observed coefficients.

B2. Bootstrapping results of LM-Parcel



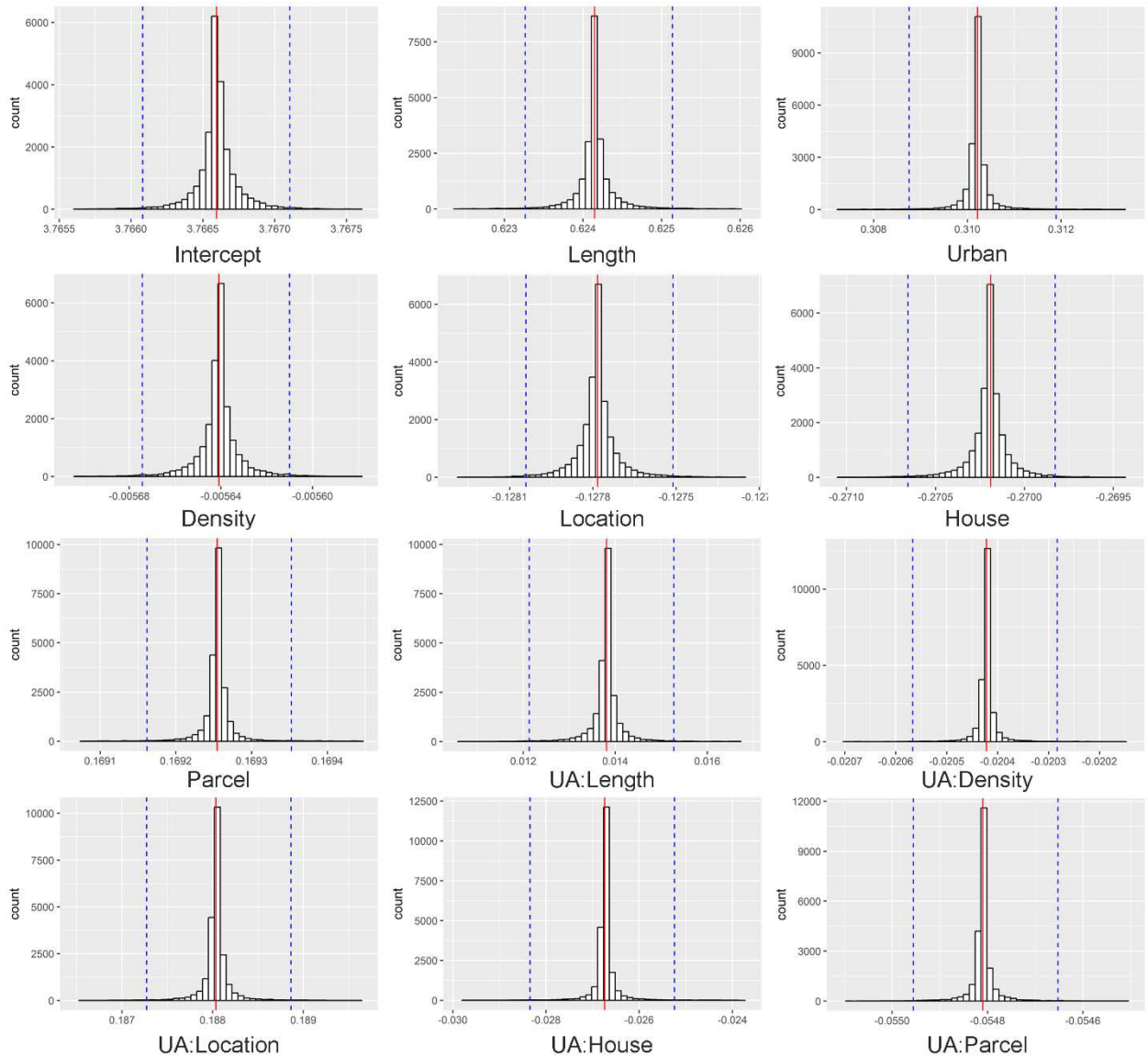
Note: the dashed vertical lines display the confidence intervals of the corresponding coefficients for a 99 percent confidence level, and the red vertical lines mark the observed coefficients.

B3. Bootstrapping results of SAR-Building



Note: the dashed vertical lines display the confidence intervals of the corresponding coefficients for a 99 percent confidence level, and the red vertical lines mark the observed coefficients.

B4. Bootstrapping results of SAR-Parcel



Note: the dashed vertical lines display the confidence intervals of the corresponding coefficients for a 99 percent confidence level, and the red vertical lines mark the observed coefficients.

REFERENCES

- Aerts, J. C., K. C. Clarke, and A. D. Keuper. 2003. Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science* 30 (3):249–261.
- Anselin, L., and S. Rey. 1991. Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* 23 (2):112–131.
- ANSI (American National Standards Institute). 1998. *Spatial Data Transfer Standard (SDTS) - Part 1. Logical specifications*. Washington, D.C.: ANSI NCITS 320-1998.
- Armstrong, M. P., N. Xiao, and D. A. Bennett. 2003. Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers* 93 (3):595–623.
- Beard, M. K., I. Leader, B. P. Battenfield, and S. B. Clapham. 1991. *NCGIA Research Initiative 7 Visualization of Spatial Data Quality*. Technical Report 91-26. Santa Barbara, CA: National Center for Geographic Information and Analysis.
- Bertin, J. 1983. *Semiology of graphics: diagrams, networks, maps*. Madison: University of Wisconsin press.
- Bhalerao, A. H., and N. M. Rajpoot. 2003. Discriminant feature selection for texture classification. In *Proceedings of the British Machine Vision Conference*, ed. R. Harvey and A. Bangham, 1–80. Durham, UK: British Machine Vision Association Press.
- Bichler, G., and S. Balchak. 2007. Address matching bias: ignorance is not bliss. *Policing: An International Journal of Police Strategies & Management* 30 (1):32–60.
- Bonner, M. R., D. Han, J. Nie, P. Rogerson, J. E. Vena, and J. L. Freudenheim. 2003. Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology* 14 (4):408–412.
- Burra, T., M. Jerrett, R. T. Burnett, and M. Anderson. 2002. Conceptual and practical issues in the detection of local disease clusters: A study of mortality in Hamilton, Ontario. *The Canadian Geographer* 46 (2):160–171.

- Cayo, M. R., and T. O. Talbot. 2003. Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics* 2: 10.
- Chow, T. E., N. Dede-Bamfo, and K. R. Dahal. 2015. Geographic disparity of positional errors and matching rate of residential addresses among geocoding solutions. *Annals of GIS* 22 (11):29–41.
- Chrisman, N. R. 1991. The error component in spatial data. In *Geographical Information Systems: Principles and Applications* Vol.1, eds. D. J. Maguire, M. F. Goodchild, and D. W. Rhind, 165–174. Harlow, UK: Longman Scientific and Technical.
- Coleman, G. B., and H. C. Andrews. 1979. Image segmentation by clustering. *Proceedings of IEEE* 67 (5):773–788.
- Cromley, R. G. 1996. A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographical Information Systems* 10 (4):405–424.
- Cromley, R. G., and G. M. Campbell. 1991. Noninferior bandwidth line simplification: algorithm and structural analysis. *Geographical Analysis* 23 (1):25–38.
- Davis, T. J., and C. P. Keller. 1997. Modelling and visualizing multiple spatial uncertainties. *Computers & Geosciences* 23 (4):397–408.
- Deitrick, S., and E. A. Wentz. 2015. Developing Implicit Uncertainty Visualization Methods Motivated by Theories in Decision Science. *Annals of the Association of American Geographers* 105 (April):1–21.
- Delignette-muller, M. L., and C. Dutang. 2015. fitdistrplus : an R package for fitting distributions. *Journal of Statistical Software* 64 (4):1–34.
- DeLuca, P. F., and P. S. Kanaroglou. 2008. Effects of alternative point pattern geocoding procedures on first and second order statistical measures. *Journal of Spatial Science* 53 (1):131–141.
- Dent, B. D. 1999. *Cartography-Thematic Map Design*. Boston: WCB McGraw-Hill.

- Drecki, I. 2002. Visualisation of uncertainty in geographical data. In *Spatial Data Quality*, eds. W. Shi, P. Fisher, and M. Goodchild, 140–159. London and New York: Taylor & Francis.
- Dutton, G. 1992. Handling positional uncertainty in spatial databases. In *Proceedings 5th International Symposium on Spatial Data Handling*, 460–469.
- Edwards, L., and E. Nelson. 2001. Visualizing data certainty: A case study using graduated circle maps. *Cartographic Perspectives* (38):19–36.
- Erkut, E., R. L. Francis, and A. Tamir. 1992. Location problems on tree networks. *Network* 22:37–54.
- Erskine, R. H., T. R. Green, J. A. Ramirez, and L. H. MacDonald. 2006. Comparison of grid-based algorithms for computing upslope contributing area. *Water Resources Research* 42 (9):1–9.
- ESRI (Environmental Systems Research Institute). 2016. *ArcGIS Desktop: release 10.5*. Redlands, California, USA.: Environmental Systems Research Institute.
- Fisher, W. D. 1958. On grouping for maximum homogeneity. *Journal of American Statistical Association* 53 (284):789–798.
- Francis, J., N. Tontisirin, S. Anantsuksomsri, J. Vink, and V. Zhong. 2015. Alternative strategies for mapping ACS estimates and error of estimation. In *Emerging Techniques in Applied Demography*, eds. M. N. Hoque and L. B. Potter, 247–273. Springer Netherlands.
- Gahegan, M., and M. Ehlers. 2000. A framework for the modelling of uncertainty between remote sensing and geographic information systems. *ISPRS Journal of Photogrammetry and Remote Sensing* 55 (3):176–188.
- Goodchild, M. F. 1991. Issues of quality and uncertainty. In *In Advances in Cartography*, ed. J. C. Müller, 113–139. London and New York: Elsevier Science Publication Ltd.
- Goodchild, M. F., B. P. Battenfield, and J. Wood. 1994. Introduction to visualizing data validity. In *Visualization in Geographical Information Systems*, eds. H. M. Hearnshaw and D. J. Unwin, 141–149. New York: John Wiley & Sons.

- Griffith, D. A. 1989. Distance calculations and errors in geographic databases. In *Accuracy of spatial databases*, eds. M. F. Goodchild and S. Gopal, 81–90. New York: Taylor & Francis.
- . 2008. Spatial autocorrelation and random effects in digitizing error. In *Spatial Uncertainty vol I. Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, eds. J. Zhang and M. F. Goodchild, 94–102. Shanghai, China: World Academic Press.
- Griffith, D. A., M. Millones, M. Vincent, D. L. Johnson, and A. Hunt. 2007. Impacts of positional error on spatial regression analysis: A case study of address locations in Syracuse, New York. *Transactions in GIS* 11 (5):655–679.
- Griffith, D. A., and D. W. Wong. 2007. Modeling population density across major US cities: A polycentric spatial regression approach. *Journal of Geographical Systems* 9 (1):53–75.
- Griffith, D. A., D. W. Wong, and Y. Chun. 2015. Uncertainty-related research issues in spatial analysis. In *Uncertainty Modelling and Quality Control for Spatial Data*, eds. J. Shi, B. Wu, and A. Stein, 1–11. London: Taylor & Francis Group/CRC Press.
- Harada, Y., and T. Shimada. 2006. Examining the impact of the precision of address geocoding on estimated density of crime locations. *Computers & Geosciences* 32 (8):1096–1107.
- Hart, T. C., and P. A. Zandbergen. 2013. Reference data and geocoding quality. Policing: *An International Journal of Police Strategies & Management* 36 (2):263–294.
- Hay, G., K. Kypri, P. Whigham, and J. Langley. 2009. Potential biases due to geocoding error in spatial analyses of official data. *Health & place* 15 (2):562–7.
- Hengl, T. 2003. Visualisation of uncertainty using the HSI colour model : computations with colours. In *Proceedings of the 7th International Conference on GeoComputation*. ed. D. Martin, 8–17. Southampton, UK: University of Southampton.
- Heuvelink, G. B. M., J. D. Brown, and E. E. van Loon. 2007. A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science* 21 (5):497–513.
- Jacquez, G. M. 2012. A research agenda: does geocoding positional error matter in health GIS studies? *Spatial and Spatio-temporal Epidemiology* 3 (1):7–16.

- Jacquez, G. M., and R. Rommel. 2009. Local indicators of geocoding accuracy (LIGA): theory and application. *International Journal of Health Geographics* 8:60.
- Jenks, G. F. 1977. Optimal Data Classification for Choropleth Maps. Occasional Paper No. 2, Department of Geography, University of Kansas, Lawrence.
- Jiang, B., F. Ormeling, and W. Kainz. 1995. Visualization support for fuzzy spatial analysis. In *ACSM/ASPRS Annual Convention and Exposition Technical Papers*, 291–300. Bethesda, MD: ACSM&ASPRS.
- Jones, R. R., C. T. DellaValle, A. R. Flory, A. Nordan, J. A. Hoppin, J. N. Hofmann, H. Chen, J. Giglierano, C. F. Lynch, L. E. Beane Freeman, G. Rushton, and M. H. Ward. 2014. Accuracy of residential geocoding in the agricultural health study. *International Journal of Health Geographics* 13:37.
- Karimi, H. A., M. Durcik, and W. Rasdorf. 2004. Evaluation of uncertainties associated with geocoding techniques. *Computer-Aided Civil and Infrastructure Engineering* 19 (3):170–185.
- Kaye, N. R., A. Hartley, and D. Hemming. 2012. Mapping the climate: Guidance on appropriate techniques to map climate variables and their uncertainty. *Geoscientific Model Development* 5 (1):245–256.
- Kohonen, T. 2001. *Self-Organizing Maps*. New York: Springer.
- Koo, H., Y. Chun, and D. A. Griffith. 2017. Optimal map classification incorporating uncertainty information. *Annals of the American Association of Geographers* 107 (3):575–590.
- . 2018. Geovisualizing attribute uncertainty of interval and ratio variables: A framework and an implementation for vector data. *Journal of Visual Languages & Computing* 44:89–96.
- Kubíček, P., and C. Šašinka. 2011. Thematic uncertainty visualization usability - Comparison of basic methods. *Annals of GIS* 17 (4):253–263.
- Lee, M., Y. Chun, and D. A. Griffith. 2017. Error propagation in spatial modeling of public health data: a simulation approach using pediatric blood lead level data for Syracuse, New York. *Environmental Geochemistry and Health*. Advance online publication. <https://doi.org/10.1007/s10653-017-0014-7>

- Leitner, M., and B. P. Buttenfield. 2000. Guidelines for the display of attribute certainty. *Cartography and Geographic Information Science* 27 (1):3–14.
- Levine, N., and K. E. Kim. 1998. The location of motor vehicle crashes in Honolulu : a methodology for geocoding intersections. *Computers, Environment and Urban Systems* 22 (6):557–576.
- Li, X. R., and Z. Zhao. 2005. Relative error measures for evaluation of estimation algorithms. In *7th International Conference on Information Fusion*: 211–218.
- Lin, M., H. C. Lucas, and G. Shmueli. 2013. Too big to fail: large samples and the p-value problem. *Information Systems Research* 24 (4):906–917.
- Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2011. *Geographical Information Systems and Science* 3rd ed. Hoboken, NJ: John Wiley & Sons.
- MacEachren, A. M. 1985. Accuracy of thematic maps / implications of choropleth symbolization. *Cartographica: The International Journal for Geographic Information and Geovisualization* 22 (1):38–58.
- . 1992. Visualizing uncertain information. *Cartographic Perspective* 13 (13):10–19.
- . 1995. *How Maps Work: Representation, Visualization, and Design*. New York: Guilford Press.
- MacEachren, A. M., C. A. Brewer, and L. W. Pickle. 1998. Visualizing georeferenced data: representing reliability of health statistics. *Environment and Planning A* 30:1547–1561.
- MacEachren, A. M., and D. DiBiase. 1991. Animated maps of aggregate data: Conceptual and practical problems. *Cartography and Geographic Information Systems* 18 (4):221–229.
- MacEachren, A. M., A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. 2005. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32 (3):139–160.
- MacEachren, A. M., R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. 2012. Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics* 18 (12):2496–2505.

- Mas, J. F., H. Puig, J. L. Palacio, and A. Sosa-López. 2004. Modelling deforestation using GIS and artificial neural networks. *Environmental Modelling and Software* 19 (5):461–471.
- Mazumdar, S., G. Rushton, B. J. Smith, D. L. Zimmerman, and K. J. Donham. 2008. Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics* 7: 13. <https://doi.org/10.1186/1476-072X-7-13>
- McGranaghan, M. 1993. A cartographic view of spatial data quality. *Cartographica* 30 (2&3):8–19.
- Mennis, J., and D. Guo. 2009. Spatial data mining and geographic knowledge discovery - An introduction. *Computers, Environment and Urban Systems* 33 (6):403–408.
- Monmonier, M. S. 1973. Analogs between class-interval selection and location-allocation models. *Cartographica: The International Journal for Geographic Information and Geovisualization* 10 (2):123–132.
- Murray, A. T., and T.-K. Shyy. 2000. Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science* 14 (7):649–667.
- Nagelkerke, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78 (3):691–692.
- Okabe, A., and K. Sugihara. 2012. *Spatial Analysis along Networks*. West Sussex: John Wiley & Sons.
- Pebesma, E. J., K. de Jong, and D. Briggs. 2007. Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. *International Journal of Geographical Information Science* 21 (5):515–527.
- Pulighe, G., V. Baiocchi, and F. Lupia. 2016. Horizontal accuracy assessment of very high resolution Google Earth images in the city of Rome, Italy. *International Journal of Digital Earth* 9 (4):342–362.
- Reyes-Aldasoro, C. C., and a. Bhalerao. 2006. The Bhattacharyya space for feature selection and its application to texture segmentation. *Pattern Recognition* 39 (5):812–826.

- Schmidt, K. S., and A. K. Skidmore. 2003. Spectral discrimination of vegetation types in a coastal wetland. *Remote Sensing of Environment* 85 (1):92–108.
- Schootman, M., D. A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo, and G. Higgs. 2007. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology* 17 (6):464–470.
- Shi, W., P. Fisher, and M. F. Goodchild. 2002. *Spatial Data Quality*. New York: Taylor & Francis.
- Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.
- Slingsby, A., J. Dykes, and J. Wood. 2011. Exploring geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics* 17 (12):2545–2554.
- Slocum, T. A., D. C. Cliburn, J. J. Feddema, and J. R. Miller. 2003. Evaluating the usability of a tool for visualizing the uncertainty of the future global water balance. *Cartography and Geographic Information Science* 30 (4):299–317.
- Slocum, T. A., R. B. McMaster, F. C. Kessler, and H. H. Howard. 2009. *Thematic Cartography and Geovisualization* 3rd ed. Upper Saddle River, NJ: Prentice Hall.
- Stacy, E. W. 1962. A generalization of the gamma distribution. *The Annals of Mathematical Statistics* 33 (3):1187–1192.
- Stephens, M. A. 1986. Tests based on EDF statistics. In *Goodness-of-Fit Techniques*, eds. R. B. D’Agostino and M. A. Stephens, 97–194. Marcel Dekker: New York.
- Strickland, M. J., C. Siffel, B. R. Gardner, A. K. Berzen, and A. Correa. 2007. Quantifying geocode location error using GIS methods. *Environmental Health* 6 (10).
- Sun, M., D. W. Wong, and B. Kronenfeld. 2014. A classification method for choropleth maps incorporating data reliability information. *The Professional Geographer* 67 (1):72–83.
- . 2017. A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. *Cartography and Geographic Information Science* 44 (3): 246–258.

- Sun, M., and D. W. S. Wong. 2010. Incorporating data quality information in mapping American Community Survey data. *Cartography and Geographic Information Science* 37 (4):285–299.
- Thomson, J., E. Hetzler, A. Maceachren, and M. Gahegan. 2005. A typology for visualizing uncertainty. In *Visualization and Data Analysis (part of the IS&T/SPIE Symposium on Electronic Imaging 2005)*. 16-20 January, San Jose, CA.
- Toregas, C., and C. ReVelle. 1972. Optimal location under time or distance constraints. *Papers of the Regional Science Association* 28 (1):131–143.
- Traun, C., and M. Loidl. 2012. Autocorrelation-Based Regioclassification – a self-calibrating classification approach for choropleth maps explicitly considering spatial autocorrelation. *International Journal of Geographical Information Science* 26 (5):923–939.
- U.S. Census Bureau. 2009. *A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know*. Washington, DC: U.S. Government Printing Office.
- Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, J. S. Colt, and P. Hartge. 2005. Positional accuracy of two methods of geocoding. *Epidemiology* 16 (4):542–547.
- Whitsel, E. A, P. M. Quibrera, R. L. Smith, D. J. Catellier, D. Liao, A. C. Henley, and G. Heiss. 2006. Accuracy of commercial geocoding: assessment and implications. *Epidemiologic perspectives & innovations* 3 (8). <https://doi.org/10.1186/1742-5573-3-8>
- Wolf, P. R., and C. D. Ghilani. 1997. *Adjustment Computations: Statistics and Least Squares in Surveying and GIS* 3rd ed.. New York: John Wiley & Sons.
- Wong, D. W., and M. Sun. 2013. Handling data quality information of survey data in GIS: A case of using the American Community Survey data. *Spatial Demography* 1 (1):3–16.
- Xiao, N., C. a. Calder, and M. P. Armstrong. 2007. Assessing the effect of attribute uncertainty on the robustness of choropleth map classification. *International Journal of Geographical Information Science* 21 (2):121–144.
- Zandbergen, P. A. 2008a. A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems* 32 (3):214–232.

- . 2008b. Positional accuracy of spatial data: Non-normal distributions and a critique of the national standard for spatial data accuracy. *Transactions in GIS* 12 (1):103–130.
- . 2011. Influence of street reference data on geocoding quality. *Geocarto International* 26 (1):35–47.
- Zandbergen, P. A., and J. W. Green. 2007. Error and bias in determining exposure potential of children at school locations using proximity-based GIS techniques. *Environmental Health Perspectives* 115 (9):1363–1370.
- Zandbergen, P. A., and T. C. Hart. 2009. Geocoding accuracy considerations in determining residency restrictions for sex offenders. *Criminal Justice Policy Review* 20 (1):62–90.
- Zandbergen, P. A., T. C. Hart, K. E. Lenzer, and M. E. Camponovo. 2012. Error propagation models to examine the effects of geocoding quality on spatial analysis of individual-level datasets. *Spatial and Spatio-temporal Epidemiology* 3 (1):69–82.
- Zhan, F. B., J. D. Brender, I. DE Lima, L. Suarez, and P. H. Langlois. 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology* 16 (11):842–849.
- Zhang, J., and M. F. Goodchild. 2002. *Uncertainty in Geographical Information*. London: Taylor & Francis.
- Zimmerman, D. L., X. Fang, S. Mazumdar, and G. Rushton. 2007. Modeling the probability distribution of positional errors incurred by residential address geocoding. *International Journal of Health Geographics* 6 (1).
- Zimmerman, D. L., and J. Li. 2010. The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics* 9 (10).
- Zimmerman, D. L., J. Li, and X. Fang. 2010. Spatial autocorrelation among automated geocoding errors and its effects on testing for disease clustering. *Statistics in Medicine* 29 (9):1025–36.

Zinszer, K., C. Jauvin, A. Verma, L. Bedard, R. Allard, K. Schwartzman, L. de Montigny, K. Charland, and D. L. Buckeridge. 2010. Residential address errors in public health surveillance data: a description and analysis of the impact on geocoding. *Spatial and Spatio-temporal Epidemiology* 1:163–8.

BIOGRAPHICAL SKETCH

Hyeongmo Koo was born in Hamyang, Gyeongsangnam-do, Korea. He earned a Bachelor of Arts in Geography Education from Seoul National University in 2005. He received a Master of Arts with a major in Geography Education in 2010 also from Seoul National University. Then he worked for Statistical Research Institute of National Statistics Office in Korea as a research associate. In January 2014, Hyeongmo moved to Texas in pursuit of his doctoral education and entered the Geospatial Information Sciences doctoral program of The University of Texas at Dallas. During his doctorate, he had the opportunity to work as a teaching and research assistant, and has published four research papers in journals. In addition, he was named a University of Texas at Dallas Esri Development Center outstanding student in 2016 and 2017, and was the winner of 2017 Korea-America Association for Geospatial and Environmental Sciences Student Paper Award.

CURRICULUM VITAE

Hyeongmo Koo

Address: 800 W. Campbell Rd, Richardson, Texas 75080-3021

Email: Hyeongmo.Koo@utdallas.edu

EDUCATION

Ph.D. (Projected completion May 2018), Geospatial Information Sciences program, School of Economic, Political and Policy Sciences, The University of Texas at Dallas

M.A. 2010, Department of Geography Education, Seoul National University, Seoul, Korea.
(Thesis: “Delineating Spatially Constrained Commuting Zones with an Improved Measurement for Functional Regionalization”).

B.A. 2005, Department of Geography Education, Seoul National University, Seoul, Korea

TEACHING/RESEARCH ASSISTANT EXPERIENCE

2014 - 2015, 2018, Teaching Assistant, School of Economic, Political and Policy Sciences, The University of Texas at Dallas.

2015 - 2017, Research Assistant, School of Economic, Political and Policy Sciences, The University of Texas at Dallas.

RESEARCH INTERESTS

Geovisualization, Spatial data uncertainty, Spatial Statistics, Spatial optimization, Geographic Information Science, and Issues related to urban economic geography

PUBLICATIONS

Journal articles

- 2018, Koo, H., Chun, Y. and Griffith, D. A., Modeling positional uncertainty acquired through street geocoding, *International Journal of Applied Geospatial Research* (in print).
- 2018, Koo, H., Chun, Y., and Griffith, D. A., Geovisualizing attribute uncertainty of interval and ratio variables: a framework and an implementation for vector data, *Journal of Visual Languages and Computing*, 44, 89-96.
- 2017, Koo, H., Chun, Y., and Griffith, D. A., Optimal map classification incorporating uncertainty information, *Annals of the American Association of Geographers*, 107(3), 575–590.
- 2014, Shin, J., Sohn, H., and Koo, H., Exploration of spatial relationship between urban regions and mountain geomorphology: review of theoretical examination and quantitative case analysis, *Journal of the Korean Cartographic Association*, 14(1), 29–47. (in Korean)
- 2013, Cho, D., Lee, S.-I., and Koo, H., Producing multilingual world maps, *Journal of the Korean Cartographic Association*, 13(1), 1-15. (in Korean)
- 2013, Lee, S.-I., Cho, D., and Koo, H., Developing a web-based multilingual geographical names service, *Journal of the Korean Cartographic Association*, 13(1), 33–46. (in Korean)
- 2012, Koo, H., Improved Hierarchical Aggregation Methods for Functional Regionalization in the Seoul Metropolitan Area, *Journal of the Korean Cartographic Association*, 12(2), 25–35.

Proceedings

- 2016, Chun, Y., H. Koo, and D. A. Griffith, A comparison of optimal map classification methods incorporating uncertainty information, in J-S Bailly, D. Griffith, and D. Josselin (eds.), *Proceedings of Spatial Accuracy 2016*. Avignon, FR: UMR 7300 ESPACE, 177–181

2015, Koo, H., Chun, Y. and Griffith, D. A., Geovisualization of attribute uncertainty, In *Proceedings of the 13th International Conference on Geocomputation*, 230–236

CONFERENCE PRESENTATIONS

2017, “Developing a spatial analysis package with ArcGIS Engine and R,” The 64th Annual North American Meetings of the Regional Science Association International, Vancouver, Canada, November 10 (Koo, H., Chun, Y. and Griffith, D. A.).

2017, “Developing a spatial clustogram and co-clustogram as new visual analytics for exploratory spatial data analysis,” The 28th International Cartographic Conference of the International Cartographic Association, Washington, D.C., July 3 (Lee, S.-I and Koo, H.).

2017, “Spatial Analysis using ArcGIS engine and R (SAAR),” The 113th Annual Meeting of the Association of American Geographers, Boston, MA, April 6 (Koo, H., Chun, Y. and Griffith, D. A.).

2016, “Modeling positional uncertainty incurred through street geocoding,” The 63rd Annual North American Meetings of the Regional Science Association International, Minneapolis, MN, November 12 (Koo, H., Chun, Y. and Griffith, D. A.).

2016, “Visualization of uncertainty in image classification with special reference to spatial autocorrelation,” The 2016 annual meeting of the Southwest Division of the Association of American Geographers, Denton, TX, October 21 (Koo, H., Chun, Y. and Griffith, D. A.).

2016, “A comparison of optimal map classification methods incorporating uncertainty information,” The 12th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Montpellier, France, July 07 (Chun, Y., Koo, H, and Griffith, D. A.).

2016, “Optimal map classification methods incorporating uncertainty information,” The 112th Annual Meeting of the Association of American Geographers, San Francisco, CA, March 30 (Koo, H., Chun, Y. and Griffith, D. A.).

2015, “Geovisualization of attribute uncertainty,” The 13th International Conference on GeoComputation, Richardson, TX, May 22 (Koo, H., Chun, Y. and Griffith, D. A.).

2015, “The classification method for defining an interval at multiple maps in spatio-temporal data,” The 111th Annual Meeting of the Association of American Geographers, Chicago, IL, April 22 (Koo, H.).

2014, “Space-time crime clusters: detection, visualization, and statistical testing using space-time kernel density estimation,” 61st Annual North American Meetings of the Regional Science Association International, Washington, D.C., November 13 (Lee, M., Koo, H., and Chun, Y.)

TECHNICAL SKILLS

GIS: ArcGIS Desktop and Server

Programming: C#, Visual Basic, Python

Statistical Analysis: R, SAS, SPSS

Remote Sensing: ERDAS Imagine, ENVI

AWARDS & HONORS

Outstanding Student Award, Esri Development Center at the University of Texas at Dallas, 2017

Student Paper Award, Korea-America Association for Geospatial and Environmental Sciences (KAGES), 2017

Outstanding Student Award, Esri Development Center at the University of Texas at Dallas, 2016