HANDLING ATTRIBUTE ACCURACY IN SPATIAL DATA USING A HEURISTIC APPROACH

by

Min Sun
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

_____	Dr. David Wong, Dissertation Director
_____	Dr. Matt Rice, Committee Member
_____	Dr. Dan Carr, Committee Member
_____	Dr. Chaowei Yang, Committee Member
_____	Dr. Anthony Stefanidis, Acting Department Chair
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: _____	Summer Semester 2014 George Mason University Fairfax, VA

Handling Attribute Accuracy in Spatial Data Using a Heuristic Approach

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Min Sun
Master of Science
George Mason University, 2010

Director: David Wong, Professor
Department of Geography and Geoinformation Science

Summer Semester 2014
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my dear parents.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. David Wong for his mentorship during my graduate studies at George Mason University.

I would like to thank all committee members for their guidance in the dissertation.

I would like to thank the U.S. Census Bureau for its partial support of the research leading to the development of this dissertation, particularly the guidance provided by Dr. Nancy K. Torrieri, the Census Bureau, ACS Office. However, I am responsible for any error in this dissertation.

I would like to thank many friends at George Mason University.

Finally, and most importantly, I would like to thank my parents. They always offer me the support I need.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

American Community Survey ..... ACS
Automated Zone Matching ..... AZM
Automatic Zoning Procedure ..... AZP
Cause-of Death ..... COD
Coefficient of Variation CV
Confidence Level CL
Class Separability ..... CS
Cumulative Distribution Function ..... CDF
Exploratory spatial data analysis ..... ESDA
Equal Interval ..... EI
Health Service Area ..... HSA
Jenks Natural Breaks ..... NB
Margin of Error ..... MOE
Modifiable Areal Unit Problem ..... MAUP
National Center for Health Statistics ..... NCHS
Probability Density Function ..... PDF
Quantile ..... Q
Standard Deviation ..... SD
Standard Error ..... SE
Surveillance, Epidemiology, and End Results ..... SEER

ABSTRACT

HANDLING ATTRIBUTE ACCURACY IN SPATIAL DATA USING A HEURISTIC APPROACH

Min Sun, Ph.D.

George Mason University, 2014

Dissertation Director: Dr. David Wong

In mapping and analyzing geographical phenomena, data are usually portrayed to be accurate without error. However, spatial data are often estimates derived from surveys, and are associated some levels of uncertainty (i.e. standard error) which make the estimates are unreliable. Ignoring uncertainty information in estimates may produce misleading results and generate spurious spatial patterns or relationships. Approaches dealing with spatial data quality have been developed decades ago, but they are mostly limited to visualize the variation of reliability, failing to incorporate data quality information in mapping and analysis. Without taking steps to address the uncertainty and its propagation in mapping and data analysis, the derived products and results may be misleading.

In this dissertation, I propose two approaches dealing with attribute uncertainty: 1) incorporating attribute error information in determining the classifications in choropleth

maps, and 2) lowering the errors in estimates by spatial aggregation to produce more reliable attribute estimates. A suite of visual analytical tools integrating different forms of dynamic data representation and user interaction interface were developed to support the implementations of the two approaches. Using these tools, users can produce data or maps with explicitly indicated quality levels.

To demonstrate the effectiveness of the proposed approaches and tools, the American Community Survey (ACS) data and national mortality data were used due to their popularity and the presence of data quality information. The proposed methods and tools are expected to be applicable to spatial data in various geographical realms, such as census, health, environment, and geopolitics, etc., as long as attribute error is included in the data product.

CHAPTER ONE INTRODUCTION

1.1 Research Problems

Surveys are used to collect social, economic, and ecological attributes of a target population in quantitative research (U.S. Census Bureau 2012a). Attributes are spatially explicit when the survey also collects location information of the observations, and each observation in the population can be geo-referenced accordingly. Therefore, such survey data may also be treated as spatial data. Spatial patterns captured by these survey data may be revealed by mapping and spatial analysis using Geographic Information System (GIS) to support decision making (e.g. Brewer and Suchan 2001; Wong and Lee 2005; Anselin et al. 2008).

However, when survey data are mapped and analyzed, they are often implicitly assumed to be accurate or with error levels that are of no significant concern. This assumption may be acceptable when data are considerably accurate, such as the 100% count data from previous decennial censuses. But when data contain substantial errors (e.g., the American Community Survey, ACS data), adopting such assumption may lead to erroneous results. Compiled information or derived products may inherit errors in data used in mapping and analytical processes (Siska and Hung 2001a; Heuvelink and Burrough 1989). Conclusions based upon the geographical patterns revealed by these inaccurate data may be misleading (Kobus et al. 2001). For example, a newspaper report claimed that Washington, DC was packed with lawyers, journalists and policy nerds

while computer scientists and biologists mostly lived beyond the Capital Beltway (de Vise 2010). Such conclusion was drawn based upon several maps compiled from the ACS data containing relatively large error in the estimated means of sampled values. The variability in sampled values can be regarded as the error of an estimate, which is usually measured by the statistic standard error. If sampled values vary tremendously, the derived estimate will contain large error, implying that the estimate is not reliable. Without considering the reliability of the mapped estimates, the observed spatial patterns and conclusions drawn from these maps are very much in doubt. In other words, if we do not consider the reliability of estimates, we will not know about the reliability of the derived information. As a result, we cannot assess the correctness of decisions based upon the data (Heuvelink and Burrough 2002).

1.2 Selected Approaches to Handle Spatial Data Quality

Fortunately, researchers realize the importance of spatial data quality (Ripley 1981; Cressie 1992; Burrough and McDonnell 1998; Goodchild 1995). Beard (1991) suggested that data users should pay attention to data quality at all stages of data processing, which include data collection, management, analysis and visualization. Usually, methods handling spatial data quality vary dramatically according to the specific quality aspect that is being studied (e.g., positional accuracy, attribute accuracy and completeness). Meanwhile, errors affecting spatial data quality are introduced from multiple sources, and different sources of error should be handled by different methods as well. Our research in this dissertation will focus exclusively on the attribute accuracy of survey data and errors affecting attribute accuracy.

Most studies on survey data quality focused on assessing error incurred during the data collection process (e.g. Su et al. 2007; Ghouse 2008), and methods assessing attribute errors are relatively mature. For instance, statistics measuring sampling error in survey data are well developed. Survey data provided by government agencies are often accompanied by measurements of attribute accuracy or at least a textual description on the methods of calculating the measurements (e.g., data gathered from the Surveillance, Epidemiology and End Results – SEER program; Griffith 1991). While statisticians have emphasized measuring data quality, Wong and Wu (1996) first advocated that data quality measurements should be stored as the attribute of features in spatial databases and treated as decision aids to assist data users. They also proposed several methods using GIS to compare data with different levels of accuracy to derive quality information. Prototype systems have been developed to handle the per-feature quality information (e.g., Gan and Shi 2002; Qiu and Hunter 2002; Devillers et al. 2005). However, these systems focus mainly on managing and storing positional accuracy information. Following this line of thought, Ghouse and Duckham (2009) introduced specific methods to store data quality information (including attribute accuracy) with features in a spatial relational database, which is efficient for spatial querying. Some GISs were developed to track the data quality information such as the number of missing values and the consistency between recorded values and value domains in a database (e.g. Unwin 1995).

With the availability of spatial data quality information, another group of researchers paid attention to communicating the quality information through mapping. MacEachren et al. (1998) pointed out that the "power of human vision to synthesize

information and recognize pattern … can mislead investigators … if data reliability is not addressed directly…" (p. 1547). Pioneering activities related to mapping data quality have started as early as 1988 (e.g., Beard et al. 1991). Nevertheless, dealing with data quality information in mapping generally is haphazard. No standard approaches or practices to include data quality information in a map have been adopted. Most solutions were only limited to simply adding error information onto existing map designs (e.g., MacEachren et al. 1998; Leitner and Buttenfield 2000). Some GISs allow users to manage, calculate and visualize spatial data quality systematically, such as the GIS Data ReViewer, which were developed by ESRI for a water company to check the quality of data recording water facilities (Esri 2004). However, such kind of systems is very few.

In addition, some researchers took into account different types of error in spatial analysis procedures (e.g., Siska and Hung 2001b). For instance, Xia (1998) considered errors in the covariates when using a Poisson spatial model to estimate lung cancer rates from count data. In Kriging, errors in attribute data can be treated as the nugget effect and presented in variogram (Carrasco 2009; Siska and Hung 2001b). On the other hand, studies outside of geosciences realm (e.g., the time series analysis using economic data conducted by Bell and Wilcox 1991) address error problems in analytical models more frequently than those in geosciences.

Review of literature leads us to conclude that research on spatial data quality related to mapping and spatial analysis is far from completed. Although map makers are encouraged to display data quality information on maps, simple display cannot reduce the error which has already been propagated from data to the resultant maps. When error is

relatively large in input data, quality of resultant maps may not be accepted. In addition, although some researchers have developed methods dealing with attribute error in their specific analytical models and datasets, no general method is regarded as the standard solution that can support most spatial analytical methods and produce more reliable results. In order to generate more reliable data products (e.g., maps), error propagation should be controlled in the mapping process. Or even better is to directly reduce the error in the original data so that mapping and spatial analysis processes can start with data with less error.

1.3 Research Objectives

Generally, two broad categories of statistical errors affect attribute accuracy in surveys: sampling error and non-sampling error (Burrough and McDonnell 1998; Sun and Wong 2010). Non-sampling error refers to the errors introduced by respondents, interviewers, coders and procedures during the data gathering and compiling processes. Sampling error exists because the sampled observations may not represent the true population well, and thus the estimate of a parameter value in the population is not highly reliable. Unlike non-sampling error which may be reduced significantly by good design and execution of survey procedure, sampling error is inherent in the data when they are collected from samples instead of all observations in the population. As a result, sampling error becomes a significant factor in affecting the estimate accuracy. Therefore, this dissertation focuses on addressing sampling error from a spatial data perspective. Sampling error can be measured by statistics (e.g., standard error), and most well-orchestrated surveys provide sampling error statistics together with estimates. If reported,

sampling error can be considered by data users. In this dissertation, I propose to address the sampling error issues with the following methods:

1) Choropleth maps, in which observations are categorized into classes, are frequently used to represent the underlying spatial patterns hidden in survey data. Spatial patterns are formed when differences between classes exhibit systematically over space. Without considering errors in estimates, map classes are usually determined using estimates only. If estimates are unreliable, an observation being assigned to one class may actually belong to other class. Accordingly, an observation may not be different from observations assigned to other classes. Thus, determining map classes without taking into account errors in estimates may create classes housing estimates that are not statistically different, and spatial patterns revealed by map classes may be fictitious. Therefore, an objective is to determine map classes that are highly separable, and the related objective is to inform map readers about the confidence levels that observations in different classes are really different.

2) Some survey data may have error too large to be used "comfortable" by users in mapping and spatial analysis. To make such survey data more usable, we may try to reduce errors in estimates. A major reason that estimates are unreliable is because of small sample sizes. Thus, to reduce errors in estimates, we may enlarge the geographical extents of observations so that the sample sizes of areal units could be raised. One possible approach to increase sample sizes is to merge areal units together, but at the expense of lowering spatial resolution of the data. Merging areal units will require a systematic spatial aggregation procedure that addresses the following issues: When

should aggregation be performed? Where the aggregation process should start and stop? How areal units should be grouped or merged, and how can the error of the merged units be calculated?

3) There is no best map classification or the best spatial aggregation result (Brewer and Pickle 2002; Unwin 1995). The most preferred solution may be a compromise produced by considering multiple objectives and selecting a choice that performs reasonably well on all measures. To reach the most preferred solution, one approach is to involve human intelligence, and produce results heuristically by evaluating the trade-offs among relevant, and sometime conflicting criteria. This heuristic process can be conducted within an exploratory data analysis environment, which is recently called visual analytics because the environment is designed to support user interaction and graphical information representation (Kohonen et al. 2001). Therefore, a set of visual analytical tools, including maps and statistical plots linked together with interactive operations and real-time calculations will be developed to support the proposed mapping process and spatial aggregation procedure in the dissertation. Within the toolset, data users can perform heuristic processes to create maps and new datasets (through spatial aggregation) with lower error levels. Demographic and public health data with sampling error information will be used to illustrate the proposed error handling approaches and the capability of the toolset.

The rest of this dissertation is structured as follows. In Chapter Two, I discuss the concept of sampling error and relevant error terminologies. The chapter also provides a review of previous studies on visualizing sampling errors, spatial aggregation methods,

and existing visual analytical techniques. In Chapter Three, I propose several methods to incorporate sampling error in determining map classes, and introduce associated visual analytical tools which can support the proposed heuristic mapping process. The spatial aggregation procedure to reduce the variability of estimates is introduced in Chapter Four, followed by the explanation of the visual analytical tools supporting the spatial aggregation procedure. In Chapter Five, I report several mapping and aggregation experiments to demonstrate the usefulness of my approaches. Finally, in the concluding chapter, I enumerate the limitations of the proposed approaches and discuss further development.

CHAPTER TWO LITERATURE REVIEW

In the past several decades, research on spatial data quality has been growing. Various terms were used in the research related to different aspects of data quality, but most of the terms are not self-explanatory. Therefore, I need to clarify these terminologies at the beginning of the study in the next (first) section. In this dissertation, one of the objectives is to incorporate data reliability information into choropleth mapping. Plenty of effort has been taken to measure error and visualize the measurement information on maps. Therefore, in the second section of this chapter, I review relevant cartographic research that deals with error information. As another objective of this dissertation is to utilize spatial aggregation to reduce error, I also review studies related to spatial aggregation. In addition, I discuss theories and fundamental concepts in existing visual analytical techniques, providing background information for the proposed visual analytical tools to support the error-handling approaches adopted in this research.

2.1 Data Quality Components and Associated Terms

The term "data quality" is rather broad and fuzzy, and other terms such as error, accuracy, uncertainty, reliability and precision are also used in the literature (e.g. Thomson et al. 2005; Howard and MacEachern 1996; Hunter and Goodchild 1996; Goodchild 2003). Generally speaking, data quality is an umbrella term referring to all quality issues that need to be considered in data collection, management, mapping and

analysis. Reliability is one aspect of data quality with a positive connotation, while uncertainty refers to the statistical nature of the data. Higher uncertainty in spatial data lowers their quality. However, different terms about data quality should not be used loosely, but correspond to different aspects of data quality. They should also reflect different sources that introduce error or may reduce the levels of data quality (MacEachren et al. 2005). Therefore, in order to clarify terms relevant to data quality, identifying different sources of lowering data quality is required.

Some research papers and spatial data standards have already sketched out different aspects of quality that need to be considered when using spatial data in different ways (e.g. Longley et al. 2001; NIST 1992; FGDC 1998). According to the spatial data transfer and metadata standards developed by the Federal Geographic Data Committee – FGDC (1998), spatial data quality may include six aspects: attribute accuracy, positional accuracy, logical consistency, completeness, lineage and cloud cover.

Completeness refers to the commission and omission in the records of features, the associated attributes and relationships (Jakobsson 2002). Logical consistency means the degree of adherence to logical rules when storing spatial data. Adherence can relate to the conceptual schema, value domains, data structure, and the correctness of topological relationship of features in datasets. Lineage describes the historical development or evolution of the data. Cloud cover, which refers to the obstruction by clouds, is often concerned about remote sensing data. Accuracy describes the difference between the observed and actual values of a particular property of a geographical object being measured (Veregin 1999).

In studying spatial data, accuracy can be further divided into positional accuracy and attribute accuracy. Positional accuracy describes the accuracy of measurements on the location of geographical objects, while attribute accuracy may refer to the accuracy of attribute measurements, or the correctness of classes to which the observations are assigned (e.g. the soil type class and land cover type class; MacEachren et al. 2005). Logical consistency may be labeled as topological accuracy when it is used in the context to describe topological relationships. Some metadata standards include temporal accuracy as well (Ostensen 2002). Temporal accuracy may refer to the correctness of timestamp associated with geographical events, the correctness of temporal sequence of a series of geographical events, and the timeliness of data collection (currency). While spatial data quality includes many aspects and the approaches to handle different aspects vary widely, this dissertation will only focus on attribute accuracy.

The accuracy of spatial data may be hampered by many factors (Figure 1). Precision generally refers to the number of digits used in recording the values (Buttenfield 1991). In the geospatial domain, precision is often related to the spatial or temporal resolution which determines the levels of detail of measurements in describing spatial objects (Veregin 1999). Improving the precision of measurements may approach capturing the reality, but there is always a limit how close one may get. The limit can be dependent upon the capability of the device used to capture and store the data. Low precision may reduce data accuracy. For example, a small river on a low resolution map may be generalized as a straight line while it actually curves around the terrain. Both the

displayed geometry and the associated attribute (e.g., length of the river) become less accurate when data are captured and represented under the lower spatial resolution.

Error is another source to reduce spatial data accuracy (Buttenfield 1991). Usually, two types of errors can be found in the data collected from survey activities: non-sampling error and sampling error. Non-sampling error refers to the errors in sampled values introduced when the information is not captured accurately. In general, this type of error is introduced due to mistakes in executing the survey procedures, such as the failure on the respondent part in providing accurate data, the failure of the interviewer to record the information correctly, and the mistakes committed by the coders in coding the survey results (Banda 2003). Therefore, non-sampling error could be minimized through employing well-trained workers and conducting quality control procedures in all phases of data collection before data are processed and released to the public (U.S. Census Bureau 2012b).

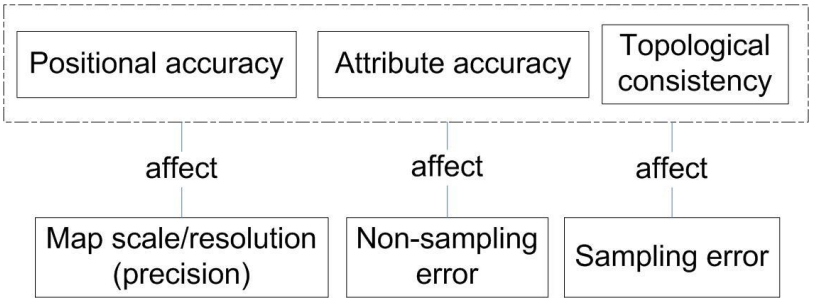

Figure 1 Relationships between selected data accuracy components and different types of error

Sampling error emerges when a sample of observations from the target population is used to infer the attributes in the population (Banda 2003). Estimates of sampled

observations are different from the population parameters. Accordingly, sampling error is related to the variation in sampled values to derive the estimates of population attributes. Due to the nature of sampling that uses only part of the population, sample data cannot provide the perfectly accurate picture of the population characteristics. Thus, sampling error cannot be removed entirely. Rather, sampling error can be reduced by various methods, including stratifying samples in proportions to the sizes of heterogeneous groups in the population or increasing the total sample size (Friedrich 2000). Theoretically, if sample size is increased to the population size, sampling error will be eliminated, as the sample and population means converge (Burt and Barber 1995). Sampling error undermines the quality of spatial data and, thus its negative impacts on spatial data processing should be controlled. I will propose several methods to handle sampling error (i.e. the variability of estimates) in this dissertation. In addition, failing to select representative samples in sampling surveys is another source that introduces differences between estimates and population parameters. This kind of difference is called sampling bias, but will not be discussed in this dissertation.

2.2 Handling Error Information on Maps

2.2.1 Significance of Attribute Accuracy in Mapping

Choropleth maps are often used to display spatial data. In choropleth maps, areal units are divided into classes according to their attribute values with selected classification methods (e.g. natural breaks, equal interval, standard deviation and quantile; Slocum 2003). Shades or colors are given to different classes, and spatial patterns emerge through the differences between classes (Wright 1938). The processes of

compiling and reading choropleth map are in no way simple even when the underlying data are 100 percent accurate, but the presence of errors in spatial data complicates these processes even further.

Xiao et al. (2007) and Sun and Wong (2010) have discussed how errors in attribute values introduce problems in compiling and reading choropleth maps. Due to the uncertainty in estimates, estimate of an observation may not be significantly different from the estimates of other observations. When producing a choropleth map, each estimate belongs to one class only to a certain extent. As shown in Figure 2, observations are assigned into three classes using the version of "natural breaks" method described in Slocum et al. (2003), which is different from the "Jenks natural breaks". The Jenks natural breaks method, instead, refers to the optimal natural breaks method developed by George Jenks (1977). Slocum et al. (2003) also described this method as the optimal method implemented with Fisher-Jenks algorithm. Observations 2 assigned to the second class has variation spanning into the first class significantly and observation 6 assigned to the third class has a reasonable probability of belonging to the second class. In other words, observations assigned to one class may not be really different from the observations in another class, or the difference between classes is not reliable.

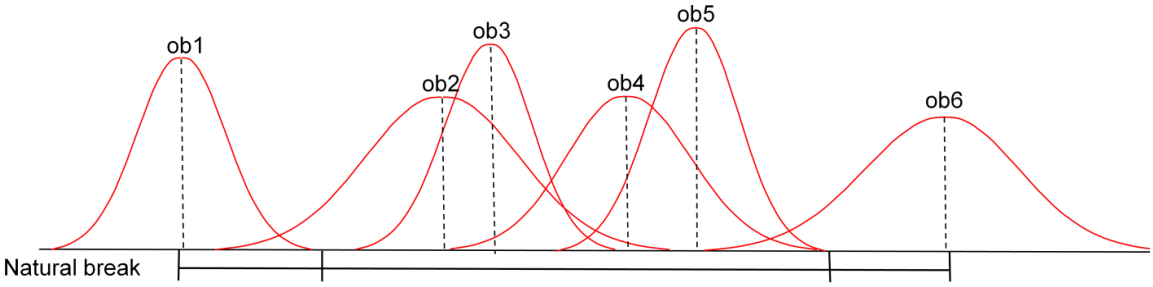

Figure 2 Estimates with normally distributed errors assigned into classes based on natural breaks method (Slocum et al. 2003)

Since some observations have significant probabilities to fall into more than one class, different maps could be generated using estimates and possible values (estimates plus variability) (Figure 3). Without any error information shown on maps, it is difficult to determine which map represents more accurate spatial pattern of the population attributes. Therefore, errors have to be recognized and presented in the processes of compiling and reading choropleth maps. If errors are not acknowledged and communicated, map readers may believe something is there but in fact nothing is of significant (MacEachren et al. 1998).

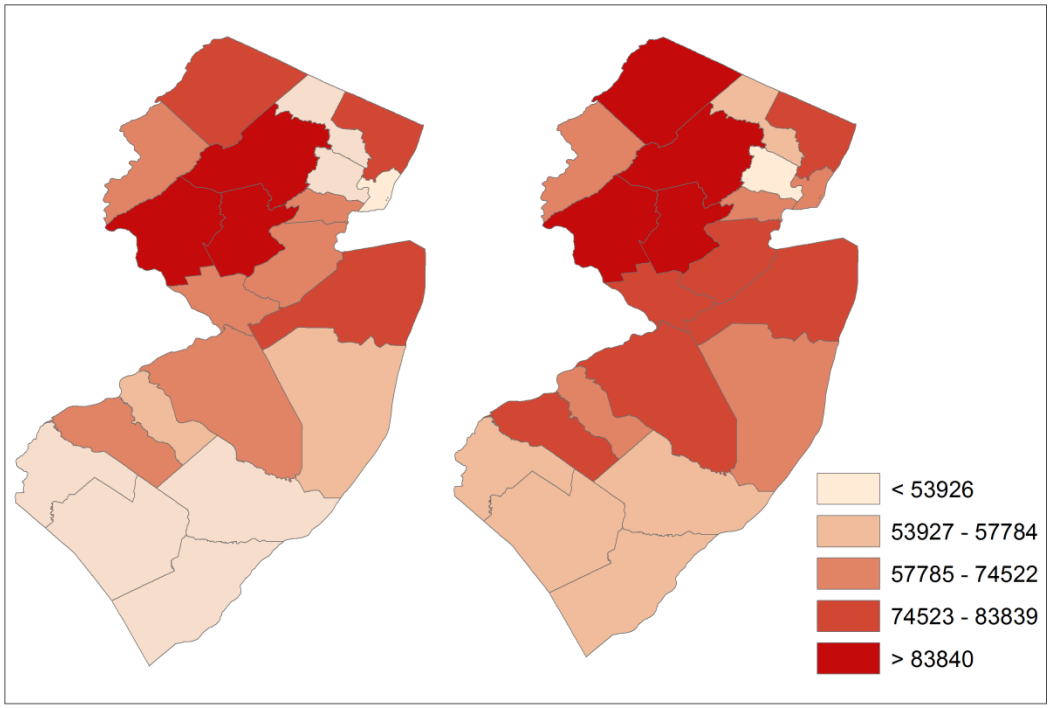

Figure 3 Maps generated using data with standard error (data source: ACS, 2011, 1-y, county, median household income, New Jersey; map on the left was rendered based on the minimum in the estimate interval at 90% confidence level; map on the right was rendered based on the maximum in the same estimate interval; the two maps share the same classes which is made based on Jenks natural break and customized by the map maker)

2.2.2 Measuring and Visualizing Error Information

Dealing with error in spatial data has a long history in GIS and cartography (Beard and Buttenfield 1991). However, how data quality information should be included on maps has not been clear. A simplistic approach, but nonetheless better than not providing data quality information at all, is to display the variability of estimate (measured by, for instance, standard error, margin of error or the coefficient of variation) on maps. This approach was adopted in the *Atlas of United States Mortality* (Pickle et al. 1996) in which box plots were used to depict the reliability of health statistics shown on the corresponding maps. Similarly, in the micromap designed by Carr and Pickle (2010),

box plots were used to display estimates and the associated variations besides the maps, which show the locations of observations included in the corresponding box plots. Observations on the map are associated with the values on the box plot with unique colors. However, when reading maps, matching the data reliability information with the corresponding estimates and spatial locations between the two displays (a map and a statistical graph) reduces the effectiveness and efficiency in assimilating the information. MacEachren et al. (1998; 2005) stated that coincident display of data and reliability information is more effective.

Two layouts for the coincident display have been proposed: 1) displaying attribute accuracy on one map accompanied by the map of the original attribute values (e.g., Leitner and Buttenfield 2000; MacEachren 1994), and 2) combining attribute and the associated accuracy information as two variables to form a bivariate map (e.g., Brewer and Pickle 2002; Olson 1981; MacEachren et al. 1998). Edwards and Nelson (2001) have compared the efficiency of visualizing attribute accuracy by separate maps and bivariate maps. The result shows that bivariate maps work significantly better than separate maps, at least when the maps are static or non-interactive.

As coincident displays are preferable, various designs of cartographic symbol have been tested to depict data reliability information (e.g., Leitner and Buttenfield 2000; MacEachren 1992). Different aspects of spatial data quality can be represented cartographically through symbols for different map elements (Thomson et al. 2005). Visualizing attribute accuracy can be achieved, for instance, by adjusting symbol fill-in (or boundary) color, color saturation, transparency, lightness, and symbol size

(Buttenfield 1991; MacEachren 1992; Burt et al. 2011; Edwards and Nelson 2001; Hengl et al. 2004), or by overlaying semitransparent cues such as hatch symbols onto the estimates of enumeration units (Xiao et al. 2007; Sun and Wong 2010). In addition, dynamic symbols, such as flicking error information on the top of estimates (Evans 1997a), were also suggested.

Among different aspects of data quality, the type of error being included in maps the most frequently refers to the basic reliability information of the original estimates. The deficiency of just showing the estimate reliability is that they may not help reflect the accuracy of information derived from data analysis, such as comparing if two estimates are significantly different or not. Some research or mapping tools have already taken into account the reliability of analytical results. In a previous version of the American FactFinder[1], it provided a web mapping tool that allowed users to compare the estimate of a selected unit with estimates of other units, determine if they are statistically different, and highlight the comparison results and significant differences on the map. Extending the idea of comparing estimates, in our previous research we developed an ArcGIS extension[2] which can help data users determine if an estimate in one class is significantly different from estimates of all areal units in other classes (Sun and Wong 2010; Wong and Sun 2013). The difference is categorized into three types: significantly different from the lower class, from the upper class, and from both lower and upper classes. However, the comparison result pertains to the estimate of a given areal unit, and thus it fails to indicate the performance of the entire class or the classification schema.

[1] http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml (last accessed on 12/21/2011)
[2] http://gesg.gmu.edu/download.htm (last accessed on 04/23/2014)

All of the above methods were used to measure the reliability of estimates in individual areal units. They cannot show the accuracy of the underlying spatial patterns presented by the maps. We cannot be certain about the spatial patterns revealed by choropleth maps partly because the classification does not consider error in data. All of the above methods cannot assess the certainty of classification or the difference among classes. On the other hand, the above methods require explicit symbolization of uncertainty, creating an additional cartographic variable that needs to be interpreted by the map reader. Depending on the map scale and information content, the added symbols may reduce map legibility. Therefore, in dealing with attribute accuracy, a more efficient approach is to incorporate accuracy information directly into the map classification to determine classes that are significantly different in values so as to improve the certainty of presented spatial patterns. This approach takes advantage of the role of classification to reduce perceptual error by generalizing multiple observations into limited numbers of classes (Dobson 1973).

Some attempts have taken into account the degree of differences between classes. For example, Jenks and Coulson (1963) noted that the number of map classes should be limited to ensure the heterogeneity between classes. The study by Stegna and Csillag (1987) went as far as determining both class breaks and the number of classes by statistically testing the difference between data-clusters (they treated a class as a data-cluster). However, all of them fell short of explicitly acknowledging that attribute data have errors, and differences among values might not be statistically significant.

The study by Xiao et al. (2007) was likely the first to explicitly evaluate the effects of data uncertainty on map classification. They defined a robustness measure as the percentage of enumeration units with the probability of belonging to a particular class greater than a pre-specified threshold. For example, a threshold of 0.8 indicates that the user will be satisfied with any enumeration unit with 80 percent of changes that it falls within the class to which it is assigned. Then, if 3/4 of all enumeration units meet or exceed this threshold, then the robustness level of the classification will be 0.75. Although robustness measures the reliability of a map classification, it operates under the premise that class breaks are given or determined *a priori*. However, none of the existing classification methods have taken account into the error information in the process of determining class breaks. As error is already in the data, without controlling the propagation of error in the classification process, we cannot improve the reliability of spatial pattern which is revealed through the chosen classification. Therefore, an approach that considering error information in the process of determining class breaks would be more useful than evaluating the overall performance of a classification method when the class breaks are given.

In our initial attempt to determine map classification incorporating errors, we place class breaks between two significantly different consecutive estimates arranging in ascending (or descending) order (Sun and Wong 2010). However, this data-driven approach suffers from at least two drawbacks. First, the algorithm compares only sequential pairs, assuming that non-sequential pairs have lower levels of significance than those sequential pairs. Then, if two sequential estimates have relatively small standard

errors, a class break may be created between them; yet estimates farther away from this class break could have large standard errors such that these estimates may not be significantly different from those estimates closer to the class break. Second, since the determination of class breaks in this method requires a predefined significance level (such as $\alpha = 0.05$), the method may not be able to create any class break if estimates have relatively large errors. Thus, a more robust approach is required to consider non-sequential pairs in determining class breaks and allow greater flexibility in seeking a balance between the degrees of difference between classes and other properties of a classification (e.g., the number of classes).

2.3 A Potential Solution to Reduce Variation on Estimates and Related Work

Incorporating sampling error in a map classification has a limitation: when error is large in estimates, observations cannot be statistically different from each other at a high confidence level. Furthermore, incorporating error in the mapping process cannot improve the quality of raw data. Using unreliable raw data in other processes such as spatial analysis could be risky. Therefore, taking a slightly different direction, the objective here is to reduce the variability of estimates in order to make data more usable.

Conceptually, sampling error of an estimate can be reduced if sample size is larger. After data are gathered and disseminated, data users cannot change the sampling procedure or the sampling error in the original data. However, by aggregating areal units together, which is analogous to increase the sampling rate, the variability of the derived estimates would be lower than the original ones. Therefore, spatial aggregation is a possible approach to accomplish our objective to reduce the variability of estimates. The

costs of using aggregation are changing the original geography (or areal unit granularity) and introduce bias to the estimates. Spielman et al. (2014) have a similar view that aggregation is one of the viable strategies to manage the uncertainty in attribute estimates.

2.3.1 The Modifiable Areal Unit Problem

In a spatial aggregation procedure, smaller areal units can be aggregated in many ways to form fewer larger areal units. Different ways to combine areal units lead to the well-known modifiable areal unit problem (MAUP) which was first formalized by Openshaw and Taylor (1979). The MAUP can be divided into two components: zoning effect and scale effect (Armhein 1995; Wong 1996). The zoning effect refers to the variability of analytical results depending on how areal units are merged; and the scale effect refers to the inconsistency of analytical results derived from data at different spatial scales or resolutions (Wong 1996).

Because of the MAUP, using highly aggregated data carries a risk: spatial pattern and relation captured in the original higher resolution data may disappear if areal units are aggregated arbitrarily (Openshaw and Rao 1995). One of the suggested solutions for the MAUP is to use heuristic algorithms involving human intelligence so as to create the most preferred aggregation among all possible schemes to suit the specific objective or undertaken data processes (Openshaw 1983; Wong 1996). Objective function is configured to measure the fitness of aggregated data to the undertaken data processes. In this dissertation, the objective function should mainly measure if the size of error in the aggregated estimate is acceptable to data users, while other criteria may be considered as

well to ensure the aggregation results are suitable for the specific application. Since spatial scale would have changed after aggregation, the potential analytical results obtained from the original data may be altered when using the aggregated data. To preserve the original results as much as possible, we should also keep the original spatial scale unchanged as much as we can during the aggregation procedure. One simple way is to involve as few areal units in aggregation as necessary, so that spatial aggregation is performed only on areal units with large estimate errors. This is different from most traditional spatial aggregation procedures that include all areal units in the study area (e.g., Openshaw and Rao 1995; Li et al. 2014).

2.3.2 Current Spatial Aggregation System

1) Existing systems

Several spatial aggregation systems have been proposed in the past. The Automatic Zoning Procedure (AZP) developed by Openshaw (1977 and 1978) for generating optimal aggregation schemes was the most popular aggregation procedure. AZP starts from a certain number of randomly aggregated zones, and then optimize the aggregation by an iterative process of recombining areal units based upon the objective function. AZP is a framework that can be used for aggregation with different purposes. For example, Openshaw and Rao (1995) implemented the AZP using Arc/Info to aggregate 1991 UK census data; Martin (1998) applied AZP to design the geography of 2001 population census in England and Wales (Martin et al. 2001), while Haining et al. (1998) developed an aggregation procedure base on the AZP to handle cancer data. In addition, Martin (2003) designed an Automated Zone Matching (AZM) system by

extending the AZP in order to merge multiple data sources under different geographies and aggregate them into larger areal units. The AZM system has been used for aggregating some environmental and health data (Cockings and Martin 2005).

However, AZP has several problems. Executing the optimization algorithm is time consuming due to the computational complexity when the number of original units or the number of desired new zones is large. Also, the result of AZP is highly dependent on the selection of seed units. To improve performance, computationally optimized aggregation algorithms have been developed, such as the simulated annealing AZP, AZP-tabu search algorithm and the simulated annealing redistricting algorithm (Openshaw 1988; Glover 1977 and 1986; Macmillan and Pierce 1994). In addition, only a few aggregation systems were developed not based upon the AZP framework (e.g. the AggRegation Tool; Guo et al. 2001). For example, Folch and Spielman (2014) defined a regionalization/aggregation algorithm which does not require the numbers of new regions to be defined before the regionalization process. They considered that the number of new regions is entirely unclear before finishing the regionalization process.

Despite the wide adoption of AZP, it cannot directly meet our objective of reducing errors in estimates. In AZP, all areal units are re-grouped into new zones, and such process is against our desire of preserving original data to the greatest extent. Also, aggregation should not start with a set of randomly selected units, but, instead, the seeds should be the units which have larger error in their estimates. As a result, we need to design a new spatial aggregation procedure for the research question.

One common problem that previous studies have not addressed sufficiently is when two or more variables are involved in defining the objective function. This problem occurs when, for instance, there is a need to compare two attribute variables. If the aggregation process is taken on the two variables independently, then the two variables may follow two different spatial configurations despite the two variables were under the same spatial configuration before aggregation. Comparison cannot be easily conducted on the variables with different geographies. The challenge is thus to keep the same geography for both variables during aggregation. A particular study by Openshaw (1983) involved two variables in aggregation when testing the zoning effect on changing the correlation analysis results. However, new areal units were generated in the random aggregation system without any constraint for the two variables in the objective function. Therefore, we need to introduce a new procedure that allows aggregation to be performed on more than one variable, but maintaining the same spatial configuration for both variables while the objective function should be defined by both variables and the levels of error in the estimates.

2) Popular aggregation criteria

The process of spatial aggregation has to be guided or constrained by one or more criteria according to specific purpose(s) of aggregating areal units. Several typical criteria for aggregation have been identified in the past (e.g. Cockings and Martin 2005 and Guo et al. 2001). For applications, such as detecting spatial pattern and correlation analysis (e.g. Openshaw and Alvanides 1999), selecting units with similar attribute characteristics is desirable to avoid too much variation to be smoothed out by aggregation. Uniform

population size across new zones is usually a criterion for generating census units (e.g. Sammons 1977; Hess et al. 1965; and Datta et al. 2012).

From a spatial data aggregation perspective, spatial contiguity is an essential criterion (Datta et al. 2012; Guo et al. 2001; Openshaw 1977). Many applications, such as rezoning survey units and political districting (Datta et al. 2012; Horn 1995; Hess et al. 1965; Briant et al. 2008), require resultant zones to be highly compacted or in relative uniform shapes, such that resultant regions may be more likely to be comparable and less biased toward certain outcomes. In practice, criteria to guide spatial aggregation procedure usually consider both spatial and thematic characteristics of the areal units (Datta et al. 2012; Hess et al. 1965, Guo et al. 2001; Sombroek and Carvalho 2000). While reducing error is the major criterion in the proposed aggregation procedure, the above criteria can be selected by data users and used to guide the determination of the way to combine areal units, so that resultant data will be suitable for the particular use purpose.

2.4 Visual Analytics

2.4.1 Potentials or Utilities of Visual Analytics

Both choropleth mapping and spatial aggregation are heuristic processes. In choropleth mapping, many objectives or criteria, such as maximizing internal class homogeneity or equalizing numbers of class members, have been suggested to guide map construction (Slocum et al. 2003). However, these objectives are often in conflict with each other. A choropleth map best for one objective may not be the best for others (Xiao and Armstrong 2006). One approach is to find the compromise on one or multiple

objectives. In our research here, making compromises will require the intelligence of map makers to determine how to trade-off conflicting objectives with lowering estimate errors as one of the objectives. For example, a map with classes that are significantly different in values may have lower performance on other objective(s). Map makers have to balance between the degrees of differences between classes and other objective(s). This experiment-based and case-dependent process is a heuristic process.

Similar to the process of creating a choropleth map that needs to consider multiple objectives, deriving zones through spatial aggregation is also driven by an objective function defined by multiple criteria (Cockings and Martin 2005). In this dissertation, aggregated zones should have error small enough to be acceptable by data users. Human intelligence needs to be involved in the entire process to determine the acceptable error level and the most preferred aggregation scheme.

Therefore, using choropoleth mapping and spatial aggregation processes to deal with sampling error share several common characteristics: 1) multiple criteria pertaining to both spatial and attribute characteristics are considered; 2) human factors are involved in evaluating the trade-offs between multiple objectives or criteria; 3) intensive computations are required to enumerate all potential options for data users to evaluate; and 4) the "best" solution is application-dependent. Proper computation tools or a system is in need to handle the computation, analysis and selection in the heuristic processes with such characteristics.

Exploratory spatial data analysis (ESDA) system, which facilitates the interaction between human and data processing through innovative graphics, maps, and efficient

computational tools, provides the potential technical support for the heuristic processes (Anselin 1994). ESDA is also called visual analytics recently due to the integration of more advanced visual techniques (Wong and Thomas 2004). ESDA system (or visual analytical system) is usually used to support a process, which: 1) does not have an unique solution; 2) engages human factor; 3) needs to consider multiple factors; 4) requires strong linkage between maps and statistical analysis; and 5) requires to combine visual and geocomputational approaches within the same environment (Gahegan 2000). All these properties correspond to the characteristics of our proposed heuristic processes dealing with sampling error in spatial data. Therefore, to support the implementations of the proposed mapping and aggregation procedures, we developed such a visual analytical system.

Koua and Kraak (2008) argued that an ESDA system (or visual analytical system) should include statistical computation capabilities, data visualization, interaction techniques, and data management if data are complex and large (Figure 4). Data visualization relates to different forms of data representation, including maps presenting geo-referenced data and statistical plots presenting attribute information. Different forms of data representation focus on different aspects of knowledge that can be derived from data (e.g., estimate distribution and error distribution). All forms of representation should be synchronized so that data analysts may recognize the underlying relationships among different aspects by their representations. In order to enabling the proposed error handling methods, the system should include the additional capability of incorporating error information in the above computational and representation components.

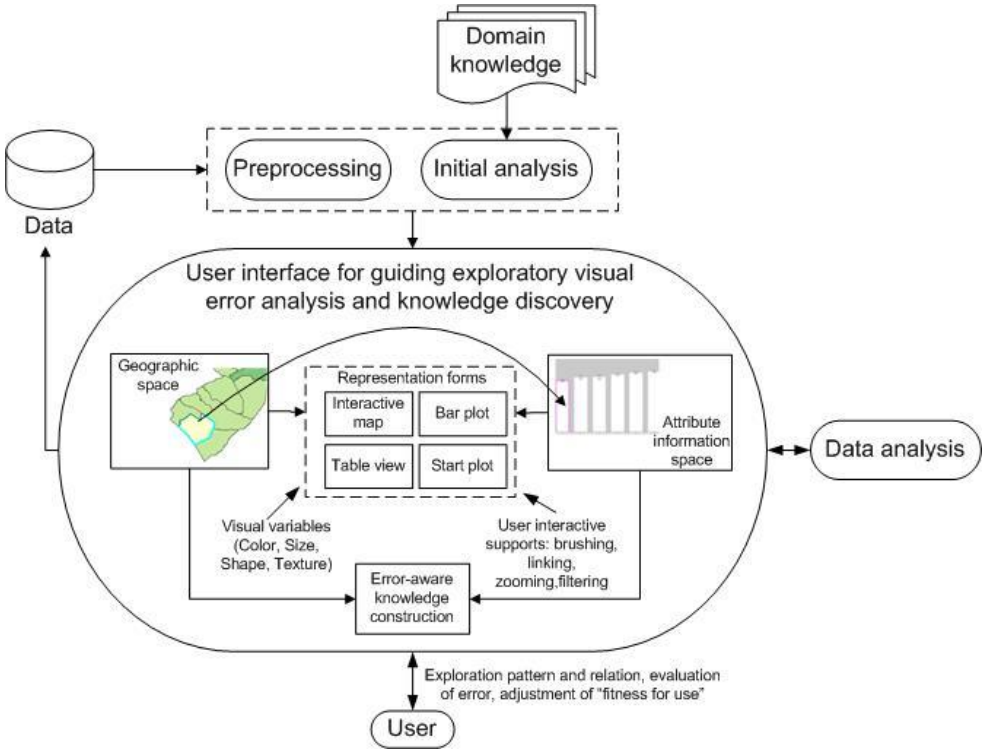

Figure 4 Components in ESDA system (Koua and Kraak 2008, p. 53)

2.4.2 Current Visual Analytical Tools

Visual analytics or related techniques have already been developed in the past for geographical analysis (e.g. Gahengan et al. 2002; MacEachren et al. 1999; Miller and Han 2001; Openshaw and Rao 1995; and Ehlschlaeger et al. 1997). Tukey (1977), Chambers (1983) and Cleveland (1994), just to name a few, have introduced statistical plotting methods to present complex data (e.g., with large number of records or in high dimensions) and associated correlations. Maps can display error information in a layer that can be turned on and off using the toggle technique and a slider bar can be used to adjust the threshold of uncertainty to alert readers (Evans 1997a; MacEachren 1993). The

conditioned choropleth map is a mapping template for exploring the potential association between multiple variables by partitioning values of a dependent variable into subsets with the control on the values of two conditioning variables (Carr et al. 2005). Slider bars were used to control class break values of two conditioning variables and maps are updated simultaneously according to any change on class breaks. GeoViz toolkit, which integrates various statistical plots and maps, was also developed for exploring relations among multiple variables (Gahegan 2002). Data brushing, which was first proposed by Monmonier (1989), was used in GeoViz to help users recognize the connections between multiple forms of information representation. Although no existing visual analytical tool or system can be directly adopted for our study, some of these concepts and components (e.g., slider bar and data brushing) can be used to develop our proposed systems.

CHAPTER THREE APPROACHES TO HANDLE ERROR IN SPATIAL DATA

Due to the unreliability of estimates, estimates in different classes may not be statistically different in a choropleth map, therefore a map classification may not reveal reliable spatial patterns (refers to Chapter 2.2.1). In order to obtain more reliable map classes, errors in estimate have to be considered in the process of determining map classes. In this chapter, I will first define a metric to measure the extent that values in different map classes are significantly different and, furthermore, introduce the approach to apply the metric in determining map classes. All methods are supported through visual analytical tools which will be introduced at the end of this chapter.

3.1 Class Separability Metric Definition

As our objective is to determine classes such that values in different classes are statistically different to the largest possible extents, we first need to determine how values between classes are compared. Although the robustness measure proposed by Xiao et al. (2007; refers to Chapter 2.2.2) evaluates the reliability of map classification, it only operates under the premise that class breaks are provided, and thus their measure is useful for evaluating the overall performance of a classification schema. In contrast, our objective is to determine the confidence levels of the statistical differences between values in any two classes. Therefore, rather than using a single measure for the entire

classification, a measure that reflects the statistical difference between estimates on two sides of a class break is needed. We label such a measure the *class separability measure*.

While such a measure may be formulated in many possible ways, our formulation starts from the basic notion of determining if the difference between any two estimates is statistically significant. In a survey context, we consider attribute value of each observation to be the mean (marked as estimate) of population values inferred from a sample. We assume that for each estimate, its standard error (SE) or alternative forms (e.g. margin of error, MOE) is provided. Although SE mainly reflects the variability of an estimate attributable to sampling error, SE may also account for other errors (e.g. non-sampling error) inevitably occurred in the survey process. The variability of estimate is assumed to conform to a normal distribution. Then, for estimates from each pair of enumeration units, we ask the question: *could the respective samples have come from the same population*? Typically, this question can be answered by testing if the estimates from the two samples are different or not. If they are not statistically different at a given significance level (such as 0.05), then the samples providing the means (estimates) for the two enumeration units may belong to a single population and the estimates from the two enumeration units should not be assigned to different classes.

Formally, we define the confidence level $CL_{i,j}$ of difference associated with the estimates of two enumeration units *i* and *j* as:

Equation 1 Confidence level of difference between two estimates

$$CL_{i,j} = \Phi\left(\frac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{SE_i^2 + SE_j^2}}\right)$$

where $\left|\bar{x}_i - \bar{x}_j\right|$ is the absolute difference between the estimates (or mean values) of the two enumeration units *i* and *j*, and SE_i and SE_j are the standard errors of the estimates graphically represented by the spreads of the two probability density functions (PDFs) in Figure 5 (upper). Equation 1 is derived from the standard *z*-test for comparing two means, assuming that the variances of the two groups are unequal and the sample distribution is close to normal. The expression in parentheses (i.e., z-test) returns the z-score (say Z_s) of the difference in observed means with respect to the normal distribution, and Φ is the function returning the probability of having a z-value of Z_s. This probability is our CL for the two units i and j that their respective estimates are different. This CL is essentially equal to the area for the acceptance region under the standard normal distribution of Z (the shaded area in Figure 5 lower).

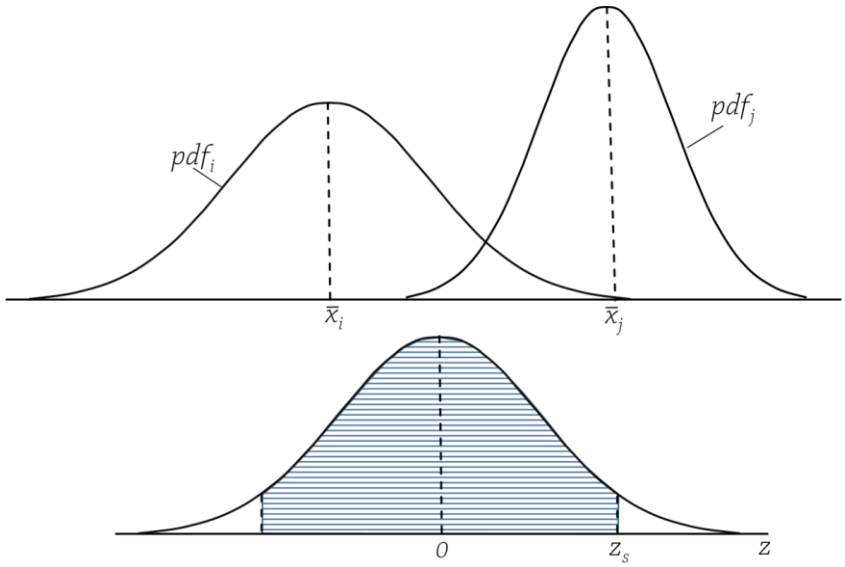

Figure 5 Overlapping probability density functions of two means (or estimates) associated with two map enumeration units (upper) and the confidence level (CL) represented by the shaded area under the standard normal distribution for a z-score less than absolute Z_s (lower).

The CL will also depend on the probability density functions (PDFs) of the estimates and their overlaps (Figure 5 upper). The curve of a PDF always extends to the two sides of the mean value and approaches to zero. Theoretically, two curves corresponding to two estimates will overlap, regardless of how different the two estimates are, and so there is a non-zero probability that the two samples actually come from a single population. Confidence level of difference is higher when overlap is less, which occurs when the SE of each PDF is smaller and/or when the centers of the two PDFs (i.e., the estimates) are further apart. In contrast, confidence level of difference is lower when the overlap is larger.

The above illustration considers only two estimates, but in practice multiple estimates are involved. Figure 6 shows a possible situation with multiple estimates with different error distributions, assigned to three classes. Ideally, we would like to be highly

confident that each estimate assigned to a class is statistically different from estimates in other classes. To determine how effective a class break is in forming differentiable groups, we have to compare all possible pairs of values. For each possible pair, a CL of statistical difference can be derived to indicate the separability between the estimate pair. Thus, we define the separability $S_{X,Y}$ between two classes X and Y as the minimum of the confidence levels among all pairwise combinations of individuals from each class:

Equation 2 Class Separability (i.e. confidence level of difference) between two classes

$$S_{X,Y} = \min_{i \in X, j \in Y} \left(CL_{i,j}\right), \qquad i \neq j,$$

For example, in Figure 6, to calculate the class separability $S_{B,C}$ between Class B and Class C, it is necessary to compute the CL between estimates X_3 and X_4 and the one between estimates X_3 and X_5. The overlap between the PDFs of the two adjacent estimates X_3 and X_4 is smaller than the overlap between estimates X_3 and X_5 because X_5 has relatively large error. That is to say that the CL between X_3 and X_5 is lower than that between the two adjacent estimates X_3 and X_4. Therefore, the separability between Class B and Class C is the CL between the non-sequential estimate pair of X_3 and X_5. Similarly, the separability between Class A and Class B is equal to the CL between X_2 and X_3, since this pair of estimates has the largest overlap (i.e., the lowest CL). Coincidentally, estimates X_2 and X_3 are a pair of sequential estimates this time. Therefore, to ensure that separability between classes are correctly determined, it is

necessary to check all pairs of enumeration units (both sequential and non-sequential) on the opposite sides of a class break and identify the pair with the lowest CL.

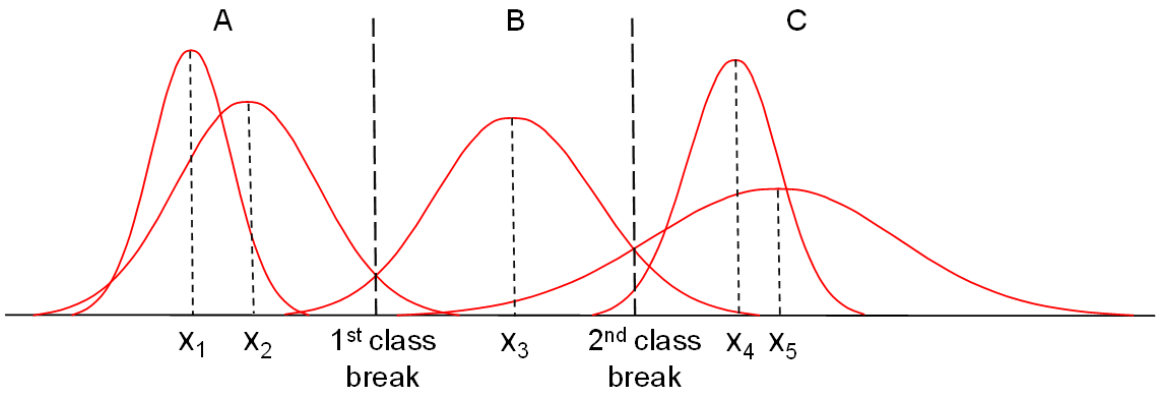

Figure 6 A sequence of estimates classified into three classes to demonstrate how to the separability between classes is determined.)

In the above example, both Classes A-B and Classes B-C are consecutive classes. However, our definition of class separability in equation 2 can be applied to any two classes, consecutive or non-consecutive. Using Figure 6 again, to identify the separability between the two non-consecutive Class A and C, the minimum CL should be determined by comparing all relevant estimate pairs, X_1 and X_4, X_1 and X_5, X_2 and X_4, and X_2 and X_5.

Note that the definition of "separability between two classes" is explicitly calculated by comparing the estimates falling into the two classes. The calculation could be slightly different for the "separability of a class break", which measures the confidence about how a class break can separate estimates on its two sides. Accordingly, the comparison should be performed on all observations on both sides of the class break.

Ideally, the observations far from the class break should have higher CL of difference as compared to those close to the break. However, with large error, observations far apart from each other may still have low CLs of difference than the ones close to each other. Since the "separability of a class break" takes all observations into account, it is, thus, more conservative than the "separability between classes" which only compares the observations within the specific classes. Both terms may be used in the content of this dissertation.

3.2 Steps in Computing Class Separability

Based on the above definition of class separability, we developed a three-step procedure to calculate this measure (Figure 7).

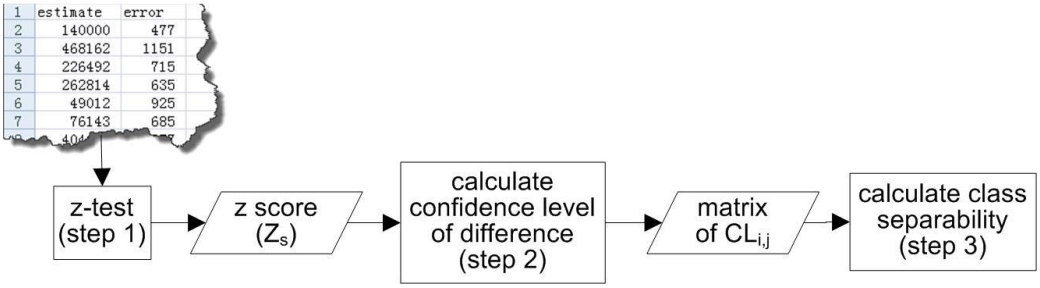

Figure 7 Three computational steps of calculating class separability associated with a class break

Step1: Calculate Z_s using two-tailed z-test for a pair of enumeration units i and j with Equation 3 *which is extracted from the part inside the parentheses in* Equation 1.

Equation 3 Z_s using two-tailed z-test

$$Z_{\mathrm{s}} = \frac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{SE_i^2 + SE_j^2}}$$

Step2: Derive the probability of having a z-value larger than Z_s and the associated CL_{ij}.

As $CL_{i,j}$ is essentially the area of the acceptance region under the standard normal distribution curve of Z, calculating $CL_{i,j}$ can be transformed into calculating the probability of having a z value between $-Z_S$ and Z_S (i.e., $P\{-Z_s \leq z \leq Z_s\}$). Cumulative distribution function (CDF), which describes the probability of estimate falling within the range $(-\infty, x]$, is used to calculate $P\{-Z_s \leq z \leq Z_s\}$. To utilize CDF, $P\{-Z_s \leq z \leq Z_s\}$ is transformed into $P\{z \leq Z_s\} - P\{z \leq -Z_s\}$. Therefore, $CL_{i,j}$ can be calculated through Equation 4.

Equation 4 Computational approach of the confidence level of difference between two estimates

$$CL_{i,j} = P\{z \leq Z_s\} - P\{z \leq -Z_s\}$$

where $P\{z \leq Z_s\}$ and $P\{z \leq -Z_s\}$ are the cumulative probabilities of having a z value equal and less than Z_S and $-Z_S$, respectively.

In addition, since two PDF curves of any observation pair must have a non-zero overlapping probability, the above comparison between enumeration units i and j will be performed for all value pairs. In essence, if we have n estimates, we need to fill an n-by-n

matrix with their corresponding CL values. Obviously, the elements along the major diagonal need not be filled or computed. Due to the symmetrical structure, only the upper triangle of the matrix is needed. According to the matrix example (Table 1, for selected counties in the state of New Jersey), the CL of the difference between the estimates of Salem and Atlantic counties is 0.51. Note that the 1.00 CLs for selected pairs were due to rounding of the decimals.

Table 1 Example of CL matrix

	Cumberland	Atlantic	Passaic	Salem	Camden	Ocean	Gloucester
Cumberland	0.00	1.00	1.00	1.00	1.00	1.00	1.00
Atlantic		0.00	0.12	0.51	1.00	1.00	1.00
Passaic			0.00	0.48	1.00	1.00	1.00
Salem				0.00	0.96	0.98	0.99
Camden					0.00	0.20	0.71
Ocean						0.00	0.63
Gloucester							0.00

Step 3: Calculate class separability

The algorithm for calculating the separability between class A and class B is described by the pseudo-code in Box 1 based upon Equation 2 that the separability is the minimum CL among all possible value pairs between the two classes.

Box 1 Algorithm of calculating the separability between class A and class B

sort all estimates in ascending order
Set x_{a1} as the smallest estimate in Class A
Set x_{a2} as the largest estimate in Class A
Set x_{b1} as the smallest estimate in Class B
Set x_{b2} as the largest estimate in Class B

```
Set x_m as an estimate in Class A
Set x_n as an estimate in Class B
set CL_min = 1
for each estimate x_a1 <= x_m <= x_a2
    for each estimate x_b1 <= x_n <= x_b2
        search CL_m,n from CL matrix if available
        if CL_m,n < CL_min then
            set CL_min = CL_m,n
        end if
    end for
end for
set separability = CL_min
```

3.3 Using the Class Separability Metric in Choropleth Mapping

3.3.1 Evaluating Classification Reliability

Errors in estimates are always ignored in choropleth mapping with conventional classification methods. However, since the knowledge of uncertainty can lead map readers to avert potential misunderstanding of spatial pattern presented by only considering estimates, it is important to evaluate uncertainty in the resultant map (Hope and Hunter 2007). Using class separability metric (Equation 2), we can evaluate the separability level for each class break determined by conventional classification methods. Meanwhile, the evaluation results should be delivered to map readers through some elements accompanied with the map of estimates. Since the separability level is associated with each class break, it is logical to attach this metric of class separability to the break values, which are shown on the legend.

To include the estimate and reliability information in the legend, the design of legend can adopt a bivariate legend template which has been used by cartographers (e.g. Pickle, et al. 1996). Traditionally, ranges of class values are shown on the right next to

the color pallets in the bivariate legend. A logical approach to indicate the separability levels is to show the separability levels (or CLs) on the left of the color pallet, but between the two color boxes corresponding to the breaks (Figure 8).

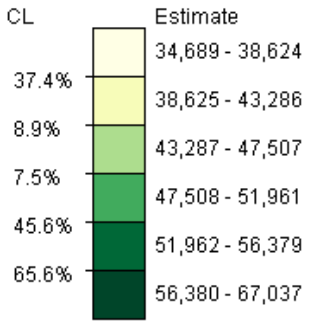

Figure 8 Design of a bivariate legend for displaying class separability

To understand the reliability information shown on the proposed legend requires map readers to have a basic understanding about the concept of class separability. Take the legend shown in Figure 7 as an example, the separability value between the second and third classes can be interpreted as the following: observations in the first and second classes are different statistically from the observations in the third, fourth and fifth classes at least 55.56 percent of the time.

3.3.2 Determining Class Breaks with Maximum Separability

In addition to evaluating the effectiveness of a class break in separating observations into different classes, the separability measure can also be used to determine class breaks that offer the highest levels of separability. Ideally, class breaks should be inserted between estimates with the highest CLs. Graphically, they correspond to the

locations with the least overlaps between PDFs. We name this classification method as "class separability".

Assuming n enumeration units in the data, separability levels can be computed for all n-1 potential class breaks, which are essentially between every pair of adjacent estimates. The separability level of each class break can be obtained by comparing all pairs of individuals on the opposite sides of that potential class break. Then desirable class breaks and classes can be determined based on these separability values.

To produce a classification scheme with *k* classes (or k-1 class breaks), the n-1 separability values are sorted from highest to lowest, and the first *k-1* potential class breaks will be used. Or, when class number k is unknown, class breaks can be determined by examining if their corresponding separability levels are higher than a minimum separability that the map maker can accept. Box 2 provides the pseudo-code that describes the algorithm to determine class breaks with the objective to maximize separability between classes.

Box 2 Algorithm to determine class breaks with maximum separability levels

Given n-1 potential class breaks which are always between two consecutive estimates
Calculate separability for all potential class breaks
Sort separability values (CLs) in descending order
a. Determining class breaks with a certain number of classes (k)
  for CL_0 to CL_{k-1}
    retrieve the corresponding estimate pair x_i and x_j
    insert a class breaks between the pair
  end for
b. Selecting class breaks according to a minimum acceptable CL (CL_{accept})
  Set CL_{accept}
  for CL_0 to CL_{n-1}
    if CL >= CL_{accept} then

```
            retrieve the corresponding estimate pair x_i and x_j
                  insert a class breaks between the pair
          else
            exit for
          end if
    end for
    return the class number k
```

One may argue that the objective of determining k classes given n values could be achieved by traditional clustering methods, such as k-mean clustering. However, most traditional clustering methods cannot take into account the reliability information of the values, which is critical in the current research context. Thus, a new method for choropleth map classification and to determine class breaks is needed.

The proposed process of selecting class breaks with sequentially lower CLs may result in a conundrum: as the number of classes increases, remaining class breaks are less and less separable. As Xiao et al. (2007) pointed out, fewer classes provide more robust classification while more classes produces less robust classification. An inherent trade-off between the class number and the separability levels between classes is apparent. Cartographers may choose to have fewer but highly separable classes, or more but less separable classes.

Another issue related to determining class breaks refers to calculating the actual values of class breaks. Break values can be selected from several options. Usually, a class break can be equal to the lowest value of the upper class, the highest value of the lower class or their average (Slocum et al. 2003). The actual break values mainly depend on how map makers want to show the ranges of classes. However, unlike the conventional

mapping practices without considering data reliability, values of class breaks in this study are crucial because, sometimes, they affect the evaluation results on the reliability of classification. However, the separability measurement we proposed is not affected by the exact class break values, because the measurement is only related to the sizes of overlapping areas between the PDFs of estimates (refers to Chapter 3.1). But the robustness measure proposed by Xiao, et al. (2007) evaluates the probability of an observations falling into a class, and therefore the reliability of classes measured by robustness may vary according to the actual break values (or the ranges of classes).

Therefore, we need to determine the precise break values so that the probabilities that observations fall into their correct classes are maximized for all. To illustrate where to find these break values, we assume two potential class breaks could be placed between the estimates of Units i and j (Figure 9 upper). While Break point 1 was placed on the intersection between the PDFs of the two estimates, the value of Break point 2 is selected randomly. If we choose Break point 2, the probability for Unit i to be assigned to the correct class will be smaller than choosing Break point 1, while the probability for Unit j will be just the opposite. Choosing a value to the right of Break point 1 will have similar, but opposite results. However, no matter choosing a value to the left or the right of Break point 1, the probabilities for Unit i and j to be assigned to the correct class must be smaller than the total probability of choosing Break point 1. Graphically, area indicated by slash lines corresponds to the total probability for Unit i and j not falling into their assigned classes and this area is always smaller than the area indicated by light grey

(Figure 9 upper). Therefore, choosing Break point 1 is likely "optimal" since it maximizes the probabilities of assigning to the correct classes.

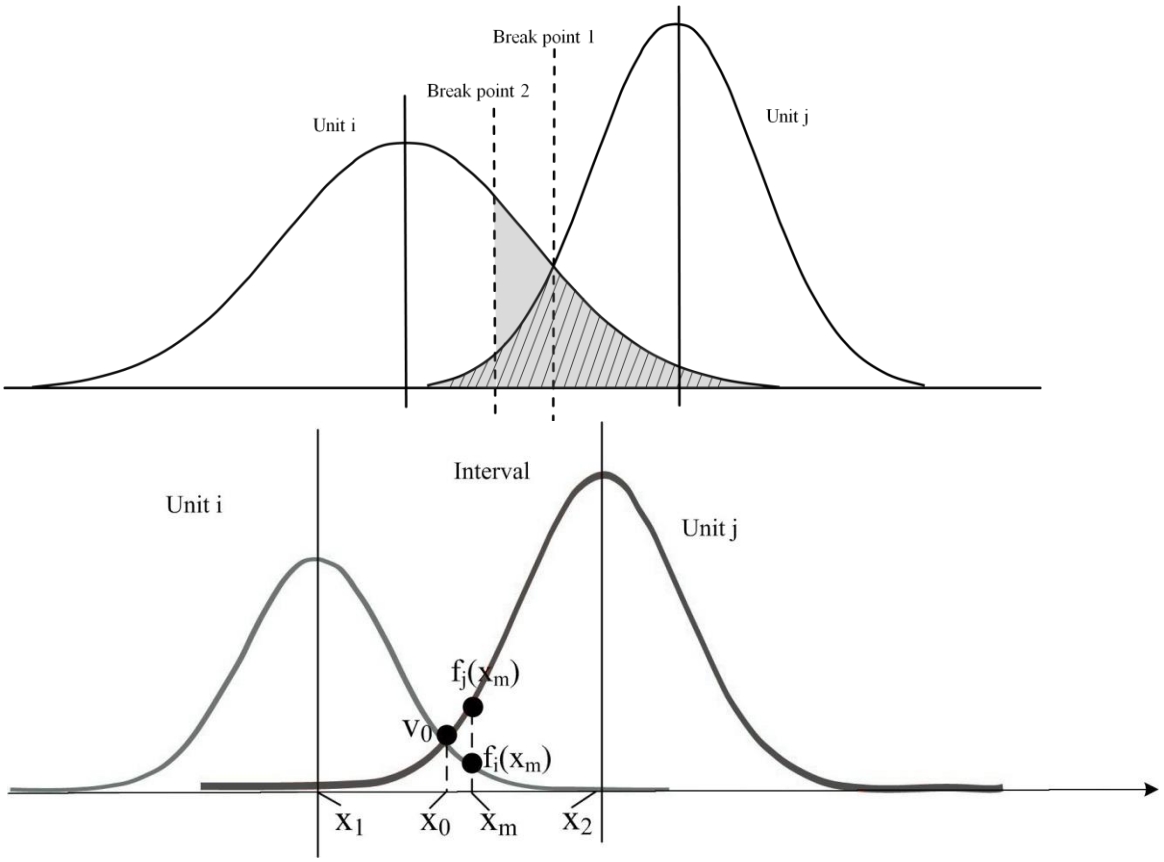

Figure 9 Possible values for a class break between two estimates of enumeration units i and j (upper: area with slash lines is determined by the Break value 1, and the area colored by light grey is associated with Break value 2; lower: v_0 is the intersection of the PDFs of the two estimates.)

Computationally, searching the intersection point is an iterative process by continuously testing a potential break value until the probabilities of that values calculated from both PDFs are the same (Box 3), i.e., $f_i(x_0) = f_j(x_0)$ (or in fact, the two probabilities are close with a difference smaller than a very small threshold). Note that, f_i

and fj are the respective PDFs of the estimates for Units i and j. We start the search from the known values (the two estimates). By setting the two variables x_1 and x_2 to the two estimates of Units i and j, then the potential break value x_m is the average of x_1 and x_2. If the probability $f_i(x_m)$ is larger than the probability $f_j(x_m)$, x_m is on the left side of x_0. As a result, x_m is set to x_1. On the contrary, if the probability $f_i(x_m)$ is less than the probability $f_j(x_m)$, x_m is on the right side of x_0, and then x_m is set to x_2. By repeatedly reassigning x_1 and x_2, according to the steps above, x_m will approach the value of intersection point x_0. When the difference between $f_i(x_m)$ and $f_j(x_m)$ is smaller than a very small threshold value, the iteration will stop. The value of the class break for units i and j is equal to x_m at the end of the iteration.

Box 3 Algorithm of calculating class break values

initial x_1 and x_2
$x_m = (x_1+x_2)/2$
while not $f_i(x_m)$ and $f_j(x_m)$ approximate the same
    if $f_i(x_m) > f_j(x_m)$ then
        $x_1 = x_m$
    else if $f_i(x_m) < f_j(x_m)$ then
        $x_2 = x_m$
    record x_m
    $x_m = (x_1+x_2)/2$
end while

3.3.3 Heuristic Mapping

As maximizing separability does not give rooms to accommodate other classification criteria, mapping using the class separability method may produce undesirable classification results. For example, in the situation in Figure 9, three classes

were made by inserting two class breaks between the first and the second observations and between the second and the third observations because the two breaks have the highest separability levels (Figure 10). As a result, two classes were created with single observation and a relatively large number of remaining observations fell into one class. Such a highly unbalanced classification is always undesirable. A more desirable map may be produced by considering additional classification criteria (e.g. the number of classes, the variation of estimates in the same class and the approximately equal numbers of observation across classes), although adopting additional criteria may result in lower class separability levels. Since there is no standards for either how reliable a classification should be or how other criteria should be considered, the most preferred mapping result may be a compromise produced by considering multiple objectives and selecting a classification schema that performs reasonably well on multiple criteria. Therefore, we propose a mapping procedure that can allow map makers to take class separability and other choropleth mapping criteria into account at the same time.

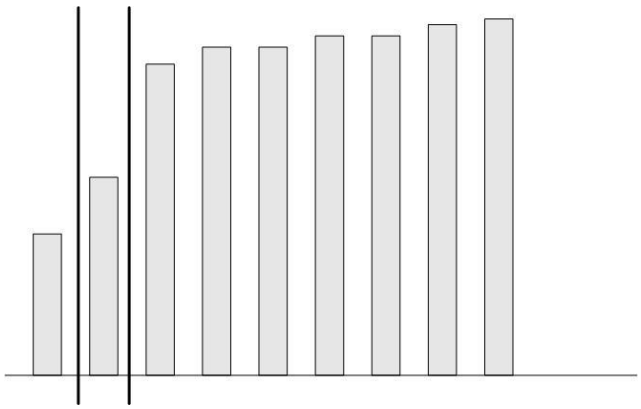

Figure 10 A highly unbalanced classification

Several criteria considered by conventional choropleth map classification methods may be used together with class separability to guide the choice of the most preferred classification.

1) Number of classes

Most of the time, the number of classes is determined by map makers arbitrarily, although there are some basic principles in determining the number of classes. Any map classification must have at least two classes, but two classes are definitely not typical in a chorolpeth map, unless it is a binary map. According to the suggestion from Brewer and Suchuan (2001), more than nine classes would not be preferable. In addition, a small dataset should not have too many classes to avoid having too few observations in some classes. Tentatively, we constrain the average minimum number of observations in each class to four, but the number can be adjusted by map makers. Accordingly, the largest number of classes is equal to the total number of observations N divided by 4 if the resultant number of classes is less than 9 (Equation 5).

Equation 5 Constraint on the number of classes

$$class\_num = \frac{N}{4} > 9 \; ? \; 9 : \frac{N}{4}$$

2) Classification robustness

Separability of the classification should be considered as a criterion. Different from the class separability measure which shows the confidence that estimates between classes are different (Equation 2), an indicator for the overall separability among all

classes is needed. We use the minimum separability among all classes. This is the most conservative indicator. The indicator is labeled as classification robustness (Equation 6).

Equation 6 Classification robustness (i.e. minimum class separability)

$$classification\ robustness = min(S_{X,Y})$$

where X and Y means any pair of neighboring classes divided by a class break, and $S_{X,Y}$ is the separability level associated with that class break.

3) Evenness of observations across classes

Another concern in classification is the distribution of observations across classes. We use evenness as a measure to evaluate the extent that observations are evenly assigned across different classes. The situation that a class has very few observations while another class has relatively large numbers of observations is not desirable (Slocum, 2003). The evenness measure is a version of standard deviation (Equation 7).

Equation 7 Evenness of observations assigned into different classes

$$evenness = \sqrt{\frac{\sum_{i=1}^{k}(n_i - \bar{n})^2}{k-1}}$$

where n_i is the number of observations in class i, $\bar{n}$ is the average number of observations across all classes, and k is the number of classes. The conventional quantile classification method has this criterion taken into account.

4) Within-class variation (dispersion)

Within-class variation is a criterion used by the natural break (or optimal natural break) method to ensure that observations with similar attribute values are assigned into the same classes (Slocum et al. 2003). Similar to class separability, within-class variation is measured for each class, and therefore we need an indicator for the overall performance of the entire classification. Descriptive statistics, such as maximum or average within-class variation among all classes could be used to summarize the overall intra-class variation. To be conservative, we choose the maximum within-class variation among all classes as the indicator and label the indicator as dispersion (Equation 8). The measure is:

Equation 8 Dispersion (i.e. the maximum within-class variation among all classes)

$$dispersion = max\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\right)$$

where x_i is the observation in one class, $\bar{x}$ is the mean of observations in the class, and n is the number of observations in the class.

Given the listed criteria and others in affecting classification, one approach to reach the most desirable classification is to involve human intelligence, and produce results heuristically by evaluating the trade-off among acceptable class separability levels and other criteria. The trade-off evaluation is actually a process of conducting experiments: evaluating several classification schemes with different numbers of classes

and different class ranges, and selecting the scheme best suited the specific mapping purpose (Monmonier 1972; Schultz 1961). Therefore, we designed the heuristic experimental mapping procedure (Figure 11).

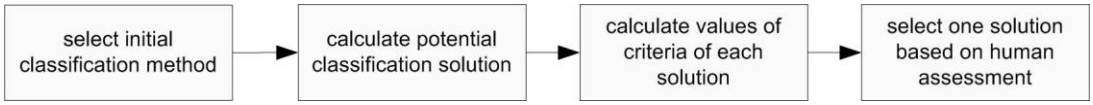

Figure 11 Major steps of heuristic mapping procedure

The heuristic mapping procedure begins with selecting an initial classification method (e.g. Jenks natural breaks, quantile, equal interval or class separability). Using the selected method, we make a classification with each of the reasonable numbers of classes. If the number of classes ranges from two to nine for the data, then we can generate eight possible classification schemes. The classification robustness and other criteria are calculated for all these classification schemes. In fact, each of the classification schemes is a combination of criterion values. Map makers should determine the acceptable levels for all criteria according to the specific mapping purpose, and then choose the 'best' classification scheme defined by the combined criteria meeting the acceptance levels. The above heuristic mapping procedure is summarized by the pseudo-code in Box 4.

Box 4 Pseudo-code for the heuristic mapping procedure

```
select initial classification method (by map maker)
calculate potential class numbers
for each class number
```

```
        determine class breaks based on the selected method
        calculate the values for the other criteria
    end for
    evaluate trade-offs (by map maker) and determine the 'best' classification
```

Sometimes, multiple classification methods may be suitable for a specific mapping purpose. For example, Jenks natural breaks and quantile are both the mapping methods designed for detecting spatial patterns (Evans 1977b). Map makers can repeat the above procedure to create possible classification schemes for the different methods and compare all the schemes to determine the "best" classification.

3.4 Visual Analytical Tools Supporting Error-aware Mapping

Both mapping processes of maximizing class separability and the heuristic mapping procedure involve the active participation of the map makers in evaluating the trade-off relationship between separability and other relevant criteria. The trade-off relationship may not be clear to the map makers at the beginning of the mapping process, but the relationship will be exposed gradually through experimentation of different classifications. Therefore, exploring the relationships between multiple criteria and selecting the most desirable classification should proceed side by side (Tukey 1977). This kind of heuristic process can be conducted within a visual analytical environment, which includes information visualization tools, real-time computational capabilities and interactive user operations (Kohonen et al. 2001).

To facilitate our proposed mapping approaches, we developed a set of visual analytical tools (Figure 12). The toolset provides an intuitive way to explore the underlying relationship among different mapping criteria and can implement all related

calculations to render classification results instantly. The toolset mainly includes: 1) a tool combining bar plot and slider bar for the mapping process of maximizing class separability, 2) a star plot to support the heuristic mapping procedure, and 3) a graphic window to render choropleth maps on the fly. Interactive operations are enabled to allow map makers to make selections or any change easily. The changes of contents in different visualization windows caused by interactive operations are synchronized in real time (Figure 13).

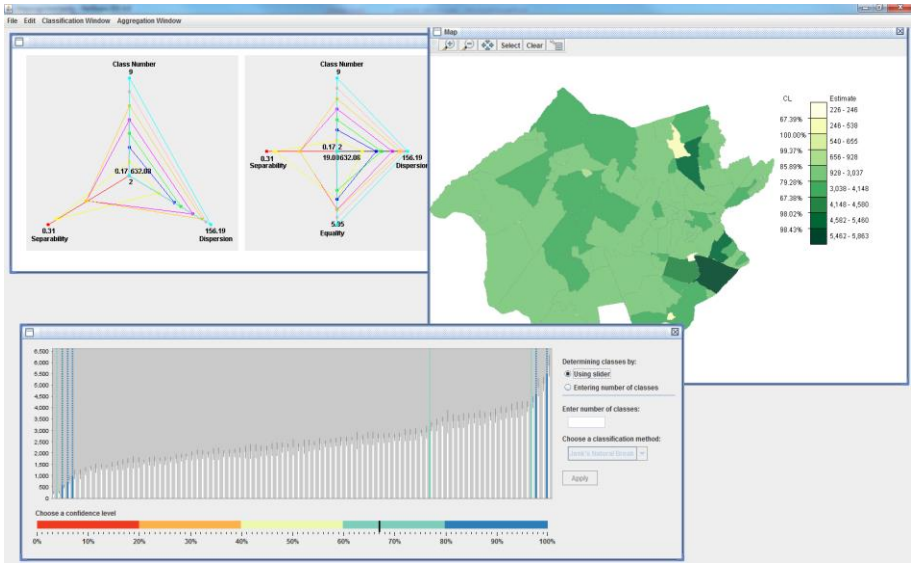

Figure 12 Screenshot of the visual analytical toolset for error-aware mapping

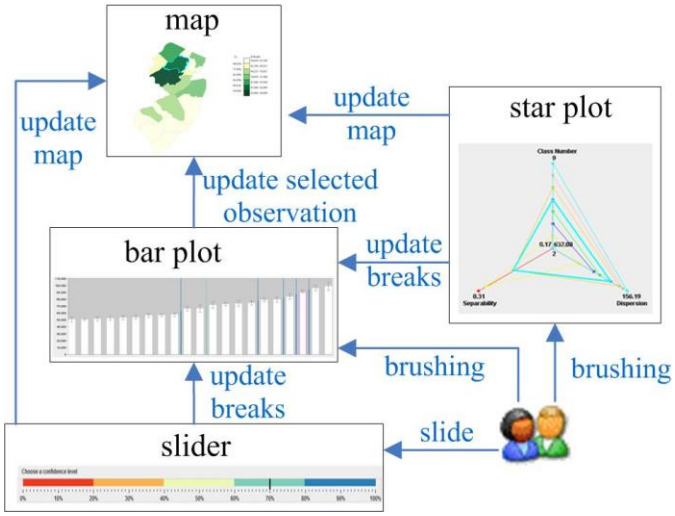

Figure 13 User interaction and linkages between different views

3.4.1 Bar plot – slider bar Mapping Tool

The bar plot-slider bar mapping tool is used to determine class breaks with the objective to maximize class separability (Figure 14). This tool consists of three major parts: 1) the bar plot to display the distribution of estimates and corresponding errors, 2) the slider bar to determine the lowest acceptable separability level, and 3) the real-time updated map to show the result of classification based on changes made on the bar plot and slider.

On the bar plot estimates of all enumeration units are shown in ascending order by vertical bars with an error bar on the top representing the error for each estimate. As mentioned in section 3.1, if a higher CL is chosen as the acceptable level, fewer CLs are larger than the acceptable level and fewer classes will be used. Conversely, if a lower CL is chosen, more class breaks with lower CLs can be used. To support the evaluation of the trade-off between class number and separability level interactively, the tool includes a

slider bar such that the cursor can be dragged to a particular separability level. This is the minimum separability among all class breaks that will be acceptable for the map classification. Class breaks will be automatically calculated when map makers move the cursor on the slider bar according to the algorithms (see Box 2 and Box 3), and placed (as vertical lines) on the corresponding locations between enumeration units on the bar plot. The height of the solid line corresponds to the actual break value and the dash portion is drawn only for a better rendering effect. Colors of the break lines are associated with different class separability levels, which are classified into five equal intervals with five unique colors on the slider bar. Red color is used for the lowest level of separability values with the purpose of warning about the low reliability of the respective classes. Any changes made on the slider bar or bar plot will trigger an automatic update of the resultant map in the map window. The separability values for class breaks are shown on the proposed legend on the map. In addition, to support queries of individual observations interactively, the bar plot and map support brushing operation.

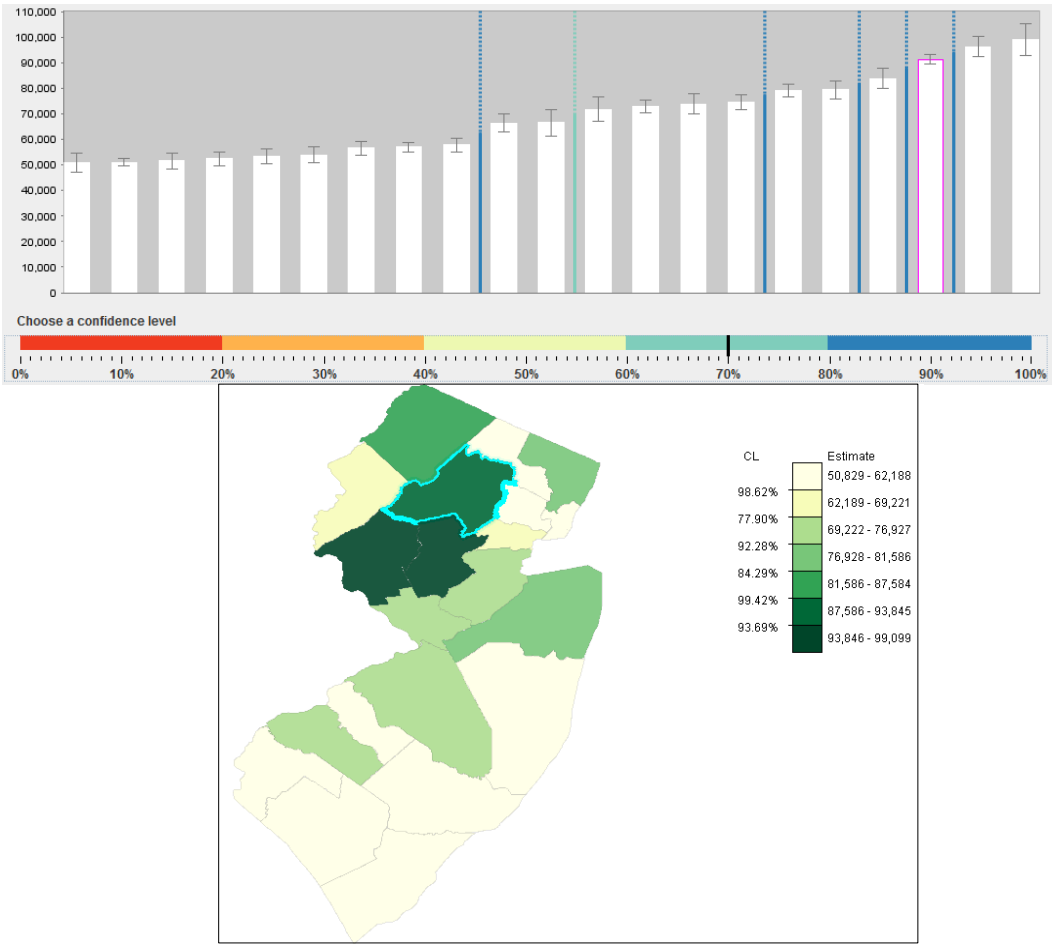

Figure 14 Screenshot of bar plot – slider mapping tool

3.4.2 Star plot Mapping tool

The proposed heuristic mapping procedure is facilitated through the tool composed of star plots and the map window (Figure 15). The star plot depicts the trade-off relationships between multiple criteria (Equation 5, Equation 6, Equation 7 and Equation 8) considered for determining map classification. Each axis of the star plot presents one selected criterion, while the number of axes depends on how many criteria are considered during the classification process. Values of selected criteria are calculated right after map makers chooses the initial classification method. All possible (and

reasonable) values of selected criteria are enumerated and represented on the star plot. Figure 15 (upper) shows two star plot examples, which include three and four classification criteria being considered, respectively. In sequence, the vertical, left and right axes of the three-axis star plot stand for the number of classes, classification robustness, and dispersion.

All possible values of each criterion are shown along an axis. In general, except the number of class, the most desirable values are at the periphery of the plot, and the most undesirable values are at the center. The number of classes is arranged from the center to the periphery in the order of two to nine. Possible combinations of different options among the criteria (i.e., different classification schemes) are linked by lines to form polygons. As the criteria have trade-off relationships, not the most desirable options for all criteria can be chosen. For instance, a map cannot have relatively similar numbers of observations in all classes and the highest classification robustness level (100%). Graphically, the classification represented by a polygon with similar distances from the geometry center to every edge has moderate performance on every criterion (e.g. the triangle highlighted in cyan in Figure 15 upper).

To assist map makers to visualize classification schemes intuitively on the map, data brushing technique is implemented to allow users to choose one classification scheme by simply clicking the corresponding polygon on the star plot. Then the choropleth map will be automatically generated based on that selected classification and shown in the map window. By comparing different classification schemes on the star plot

and associated choropleth maps side by side, map makers can decide the most preferred map for their mapping purpose.

Figure 15 Screenshot of star plot mapping tool

In summary, this chapter described a separability metric for measuring the significance of differences between map classes, the methods to use the metric in choropleth mapping and the visual analytical tools for supporting the proposed

choropleth mapping approaches which incorporate class separability. Making choropleth maps with the proposed methods will be tested and evaluated in Chapter 5 using real world data. In the next chapter, rather than handling attribute accuracy in the mapping process, we will emphasize on reducing errors in estimates from the raw data to produce more reliable data which can be used for more GIS processes, including spatial analysis.

CHAPTER FOUR REDUCING ESTIMATE UNCERTAINTY BY SPATIAL AGGREGATION

The reliability of results in mapping or spatial analysis highly relies on the quality of raw data. Taken the proposed mapping approach of maximizing the separability among classes as an example, if estimates have large error, producing highly separable map classes is difficult. Even if class separability is maximized in determining class breaks, separability levels between classes could still be very low. In other words, data with large error are difficult to use. Therefore, reducing errors in estimates is needed to make data more usable. As discussed in Chapter 2.3, variability of estimates can be reduced if sample sizes increase, and spatial aggregation is a data manipulation method to increase the reliability of estimates through merging areal units of relatively small sample sizes together. Accordingly, we need to develop such a spatial aggregation procedure, and the objective of spatial aggregation in this study is to create new areal units with estimates of improved reliability.

While spatial aggregation reduces standard errors of estimates, it lowers the spatial resolution of data and introduces bias to estimates of the aggregated units, as the new estimates of the aggregated units are different from the original estimates. Thus, the underlying spatial variation in the original data is changed. Data users need to realize the cost of bias introduced to the estimates in addition to the cost of lowering spatial resolution when they consider using spatial aggregation to reduce the error in data.

Therefore, a possible constraint that can be included in the aggregation procedure is to control the magnitude of bias that may introduce.

In order to minimize changing the spatial variation in the original data, we should only aggregate enumeration units with a reliability level too low to be acceptable to the data users. This condition separates our spatial aggregation procedure from most existing spatial aggregation procedures, which usually involve changing the boundaries of most, if not all enumeration units in the study areas (e.g. Openshaw 1983; Martin 2003; and Datta et al. 2012). One of the primary tasks in developing the aggregation procedure is to identify units that need to be aggregated. Similar to the developments of existing spatial aggregation procedures, our proposed aggregation process also faces many challenges, one of which is related to the MAUP: different combinations of areal units will produce different results (Openshaw and Taylor 1979; Wong and Thomas 2004). Openshaw (1983) strongly argued that the most appropriate response to this challenge is to design purpose-specific (or application-specific) aggregation systems. However, the "best" aggregation solution for the particular purpose may encounter complex trade-offs between competing constraints or objectives (Martin 2003). Therefore, the second primary task of this study is to develop a flexible procedure that can accommodate such complex trade-offs in the evaluation process and incorporate the main objective of reducing uncertainty in estimates.

Performing spatial aggregation based on the characteristics of one variable is relatively simple, but the resultant data may not be too useful, as many spatial data analytical procedures are multivariate in nature, such as detecting relationships among

variables, or comparing changes over time. Therefore, data produced from spatial aggregation for multiple variables should match to the same geography. The third challenge of designing our aggregation procedure is to include more than one variable in the spatial aggregation process. To overcome these challenges and fulfill the primary tasks mentioned above, we propose a heuristic spatial aggregation procedure, which integrates human inputs in the design of aggregated zones.

4.1 A Heuristic Aggregation Procedure

Assuming that at most two variables can be handled at the same time, we propose a four-step aggregation procedure (Figure 16). Each step refers to one major task, and each task includes some sub-processes.

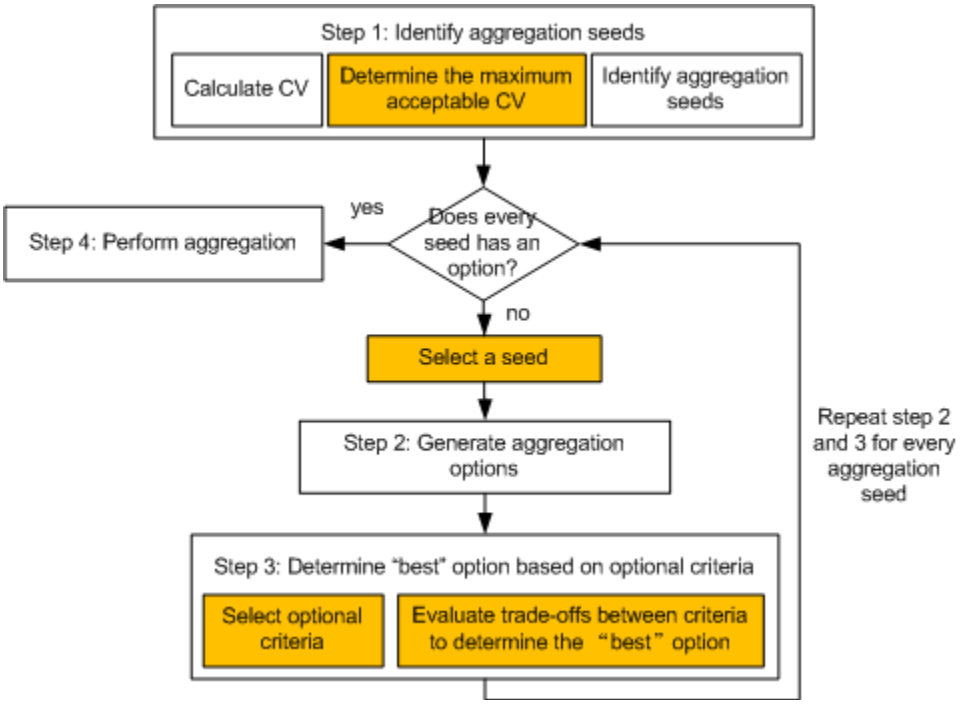

Figure 16 Steps of spatial aggregation procedure for reducing the variability of estimate (Orange rectangles refer to operations to be completed by data users)

Step 1: Identify aggregation seeds

The procedure begins with identifying a set of observations with relatively large errors that are not acceptable to data users. We need a metric to measure the level of errors in estimates. Values of some error measures, such as SE and MOE, are influenced by the absolute values of estimate. On the contrary, the coefficient of variation (CV, Equation 9) is independent of the absolute value of an estimate. CV is defined as

Equation 9 Calculation of CV

$$CV = \frac{\sigma}{\mu}$$

where μ is the estimate (or mean) and σ refers to the standard error.

In order to compare errors among observations that may have dramatically different estimates in the study area (e.g., income levels in midtown Manhattan versus Harlem, both are on the island of Manhattan, NY), CV is chosen as the measure here. Large CV implies that the estimate is not reliable and should not be used or be used with great caution. As a result, spatial aggregation could be performed on those enumeration units with such large CVs.

Data users can define an acceptable CV value as the threshold. Observations with CV larger than the threshold should be merged with other units. These observations are labeled as aggregation seeds. If aggregation is performed concurrently on two variables, data users need to provide thresholds for both variables and seeds are determined by

evaluating the CV levels of both variables. Seeds will be units having any one CV value larger than its threshold (Equation 10).

Equation 10 Aggregation seed

$$seeds \in \{CV_{i1} > CV_{accepted} \cup CV_{i2} > CV_{accepted}\}$$

For each aggregation seed, the next step is to search its neighboring units to identify the candidates that could be used to merge with the seed.

Step 2: Identify aggregation candidates/options

There are many ways of merging neighboring units to a seed. Each way is labeled as one option and the areal units in the option as candidates. An option may include a single unit or multiple units. Areal units become to the candidates to be included in aggregation option if 1) these units are contiguous to the aggregation seed; and 2) merging these units and the seed can lower the CV of the new unit below the threshold level set by data users. Accordingly, the objective function for searching aggregation candidates is actually minimizing the error for the potential zone by merging the contiguous units and the seed. In most existing procedures (e.g. Martin 2003), the searching process stops when the objective function cannot be improved significantly any more. In other words, spatial aggregation result should be close to optimal. Unlike those procedures, we do not need to search for candidates to produce the minimum variability of the new estimate, since such estimates with minimum variability may perform poorly on the other criteria. Most data users usually can tolerate a certain level on error in the

data and such acceptable error level is likely dependent on the particular application. Accordingly, areal units may be considered as candidates to form an aggregation option as long as the variability of new estimate is acceptable to the data user.

In addition, when units are merged, the variability of an estimate for the new unit has to be estimated or approximated, but for the original estimates the variability is derived with the original sample observations. The approximated estimate variability is usually different from the one directly derived with original observations. King et al. (2011) have tested such difference using ACS data. The results show that the difference becomes significant when aggregation involves more than four units. Therefore, spatial aggregation, for our purpose, should involve as fewer areal units as possible.

The algorithm searching for aggregation candidates/options is described in pseudo-code in Box 5 and shown in Figure 17. The first order neighbors of the seed are first considered to form the single-unit option. As long as we can identify options among the first order neighbors, the search process will stop. We expect the number of aggregation options being identified is more than two, so that data users can have multiple choices of option when they exercise their intelligence in evaluating the trade-offs among aggregation criteria. Consequently, if less than two single-unit options can be identified, then the number of areal units (candidates) to be included in aggregation option will increase to two (excluding the seed) and the process to search for option will start again. The two-unit options could be formed by candidates within the first order neighbors or by combining first and second order neighbors. The number of units

involved in aggregation will keep on growing until the search process results in more than one aggregation options.

In addition, when the aggregation is taken on two variables, errors of the new units need to be compared with the corresponding thresholds for both variables. The candidates to be merged with the seed must produce errors lower than the thresholds for both variables. The differences of the searching algorithm between involving one or two variables are displayed in Italic in Box 5.

Box 5 Algorithm of searching aggregation candidates and options

set unit with unacceptable CV as u (i.e. the seed) and option pool as l_{cdt} (Note that, option can only include units not being used before.)
generate a list (l_n) of first order neighbors of u
for n in l_n
  group n (the candidate) with u as an option, calculate the potential new CV (CV_{new})
  if CV_{new} is smaller than $CV_{accepted}$ *(for both variables)*
    store the option in l_{cdt}
  end if
end for
while the size of l_{cdt} is smaller than 2
  copy l_{cdt} as $l_{cdt2\ and}$ clean l_{cdt}
  for each old option (c) in l_{cdt2}
    generate a list (l_n) of first order neighborhoods for c
    for n in l_n
      group n with c as an option, calculate new CV (CV_{new})
      if CV_{new} is smaller than $CV_{accepted}$ *(for both variables)* and not contain used unit
        store the candidate in l_{cdt}
      end if
    end for
  end for
end while

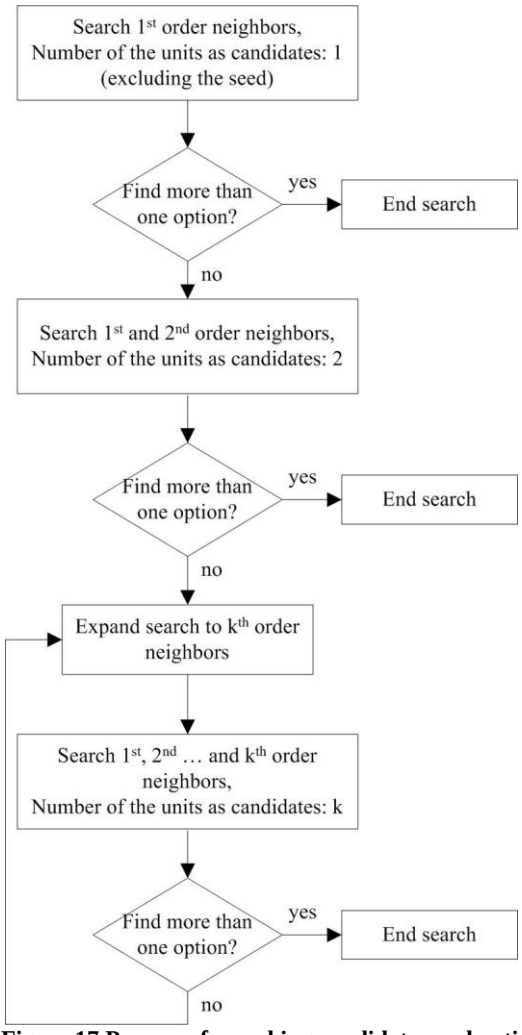

Figure 17 Process of searching candidates and options for a seed

Step 3: Determine the "most preferable" aggregation option based on optional criteria

After identifying the options, users should determine the 'best' option (i.e. a combination of candidate units) to merge with the corresponding seed. Beside the two default criteria (spatial contiguity and lower error in the new areal unit), data users may include additional criteria that capture the spatial and attribute characteristics of

enumeration units, such as the compactness of the new shape and attribute similarity, to guide the final decision. Note that, measures for additional criteria have been calculated for every option as soon as the option was identified. Users can consider the trade-offs between the reliability of the aggregated estimates and additional criteria. By considering these criteria, spatial aggregation may produce resultant data more suitable for the particular application. However, whether including these criteria in the final step of merging the candidate(s) with the seed is optional and dependent on the specific purpose of using the aggregated data. Sometimes, a potential new unit with more reliable estimate may score poorly on the other criteria. The "best" option is likely the one that has acceptable variability for the new estimate and moderate values on the other constraining criteria. In addition, when the aggregation involves two variables, the "best" option should be the one which satisfies all selected criteria in both variables.

In this study, we offer three optional criteria, which have been widely used in aggregation studies of socioeconomic data. They are the compactness of areal units, thematic similarity and spatial hierarchy. Theoretically, optional criteria are not limited to these three, but can be defined by data users according to the needs of their particular purposes. We will explain these optional criteria in section 4.2. After determining the "best" aggregation option for the current seed, steps 2 and 3 are repeated for each seed.

Step 4: Perform aggregation

After the "best" aggregation option is identified by the user for each seed, features of all involved units are merged, while a new estimate and its associated error are derived and stored in the attribute fields. However, the new areal unit, which is an aggregate of

smaller original units, cannot be labeled in the same way as in the original data any more. Taking the census tract level data as an example, the new areal unit could not be labeled as “census tract” any more, but “pseudo-census tract” is more appropriate.

4.2 Aggregation Criteria

As mentioned in section 4.1, two default criteria are used to constrain the search for aggregation candidates and options, and several optional criteria can be used to guide users to select the most preferable combination of units to be merged. Here, we explain both the default and optional criteria in details.

1) Spatial contiguity (default)

As used by many previous studies, spatial contiguity is the fundamental criterion for all spatial aggregation procedures (e.g. Ralphs and Ang 2009; Shirabe 2009; Horn 1995). This criterion requires that all areal units involved in the aggregation are either adjacent or indirectly linked to the seed through other units involved in the merging process. In this study, as merging contiguous polygons should create a relatively compact new polygon, we use the rook’s contiguity which requires two polygons to share an edge, not a point. In Figure 18, if unit 0 is an aggregation seed, polygons in dark grey are the first-order neighbors and the second-order neighbors are highlighted with light grey.

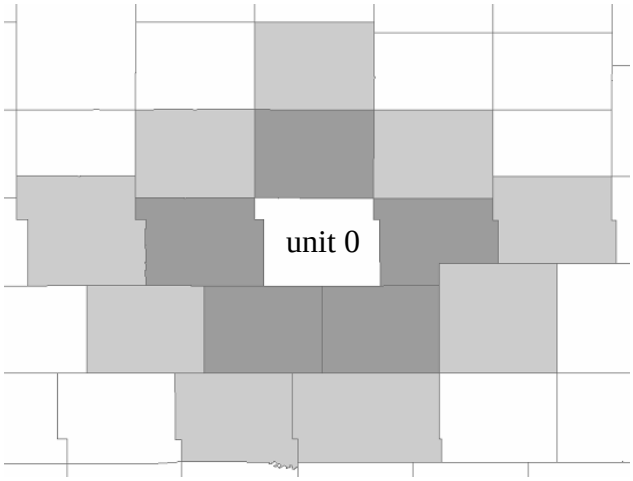

Figure 18 First-order and second-order neighbors of unit 0

2) Variability of the estimate (default)

According to the objective of the proposed spatial aggregation procedure, variability or error of the new estimate for the aggregated zone must be lower than the user-defined threshold.

3) Compactness (optional)

Compactness is a measurement to describe the geometric shape of an areal unit. This type of measures plays an important role in the context of using spatial data (Vanegas et al. 2010). The derived new zone should have a relatively compact shape (a circle is the most compact shape; Li et al. 2013). Compactness can be measured by the ratio of area to perimeter (Equation 11) (Frolov 1975; MacEachren 1985; Wong 2003). Higher value means that the geometric shape is more compact.

Equation 11 Calculation of shape compactness

$$compactness = \frac{area}{perimeter}$$

4) Thematic similarity (optional)

Thematic similarity refers to the similarity among estimates. Spatial aggregation is actually a process of smoothing the spatial variation among attribute values. To avoid smoothing the variation too much due to aggregation, the merged zone should comprise of units with high local spatial autocorrelation (Guo et al. 2001). For example, in the task of aggregating areal units with different levels of household income, it would be preferable to merge the units with similar income levels rather than merging the units with extremely high and low income levels together. Or merging areas dominated by people speaking Indonesian with areas of people speaking Malay is better than merging with the areas of people speaking French when the data are used to study the distribution of world's language, since both Indonesian and Malay belong to the Austronesian language.

Instead of simply taking the difference between the attribute values, t-test can be used to examine if the attribute values are significantly different before their corresponding units are merged. Formally, t-value is the part inside the bracket of Equation 12, and will return the t-statistic of the difference between estimates. Then the Φ function returns the probability that the estimates of two areal units are different. Accordingly, 1 minus this probability gives the probability that the two areal units are not statistically different (i.e. similarity). Similarity should be evaluated for all pairs of

candidate estimates involved in the aggregation option (including the seed), and the average of all similarity values is the value for the thematic similarity criterion.

Equation 12 Calculation of thematic similarity

$$similarity = 1 - \Phi\left(\frac{\left|\bar{x}_i - \bar{x}_j\right|}{\sqrt{SE_i^2 + SE_j^2}}\right)$$

5) Spatial hierarchy

Any areal unit should be a member of one or more administrative, political or statistical system. To determine if an areal unit should be a candidate for a seed, their memberships in these systems should be considered. For example, a census tract is more reasonable to be merged with other tracts within the same county than with tracts in another county. Thus, both the candidate(s) and the seed belonging to the same unit in the upper hierarchy of a system would be preferable. The boundary of the unit in the upper level of the geographical hierarchy is labeled as the constraining polygon/boundary.

Instead of using a simple binary yes-no variable to determine if areal units are within the same constraining polygon of an upper hierarchical level of the spatial system, we define the 'intersection ratio' to evaluate the spatial relationship between the candidate units, the seed and the polygon in which they share the membership. Given that an aggregated zone is formed by merging candidate(s)and the seed, the intersection ratio, then is the ratio of the area that the new unit falling within the constraining polygon to the total area of the new unit (Equation 13).

Spatial intersection analysis is performed to obtain the subarea of the new unit falling inside the constraining polygon. If the spatial intersection analysis is based on the 9-intersection model (Clementini 1993), the possible value calculated from Equation 13 ranges from 0 to 1. Value 1 means the new unit and the constraining polygon overlap perfectly. On the other hand, value 0 reflects the situation that the new unit is completely within the constraining polygon. Therefore, both the value 0 and 1 indicate all candidate units and the seed are within the same constraining polygon, and this is desirable. If some candidate units in the option are not completely within the constraining polygon, then the equation returns a value between 0 and 1. To make the values easier to explain, we can replace value 0 with 1, so that larger value is more desirable for this criterion.

Equation 13 Calculation of intersection ratio (spatial hierarchy)

$$intersection\ ratio = \frac{area\ within\ constraining\ polygon}{total\ area\ of\ new\ unit}$$

6) Aggregation bias

In order to avoid introducing significant aggregation bias, i.e. the difference between the values in the original areal units and the aggregated areal unit, an optional criterion can be used to evaluate the bias. Aggregation bias can be calculated by comparing the new estimate to all the original estimates which were used to derive the estimate of new areal unit, and then summing up all the differences (Equation 14). Theoretically, bias should be minimized, but data users can set their own thresholds

above which the bias is not acceptable and then aggregation should not proceed. Clearly, these threshold values are dependent upon particular cases.

Equation 14 Calculation of aggregation bias

$$bias = \sum_i (x_{new} - x_i)^2$$

Where x_{new} is the estimate of the merged areal unit generated from aggregation, x_i is the estimate of the original areal unit i which was merged with the seed to generate the new areal unit.

4.3 Calculation on Derived Estimate and Variation

New estimate and associated variability or error of the new unit have to be determined, but their calculations depend on the types of measurement, such as counts (e.g. population size), ratios (e.g. ratio of females living alone to males living alone), and percentages (e.g. proportion of single person households that are female). The U.S. Census Bureau has documented how the estimate and error can be derived (U.S. Census Bureau 2008). For example, the estimated count of the new unit is the sum of the counts of all units to be merged, and the new error (measured either by the margin of error or standard error) is calculated using Equation 15 (U.S. Census Bureau 2008). More calculation methods need to be developed for the measurements like median.

Equation 15 Calculation of the estimate and error for the new unit

$$MOE_{agg} = \pm\sqrt{\sum {MOE_i}^2} \text{ or } SE_{agg} = \frac{\pm\sqrt{\sum (SE_i * 1.645)^2}}{1.645}$$

where MOE_i and SE_i are the margin of error and standard error, respectively, of the i^{th} unit to be merged. As already mentioned in Chapter 4.1, the new MOE is just an estimate, not the accurate MOE as if calculated from all the sample observations.

4.4 Visual Analytical Tools Supporting the Heuristic Aggregation Procedure

Similar to the mapping approaches discussed in chapter 3.3, the proposed aggregation procedure is also characterized as a heuristic process. Heuristic is preferable because the procedure: 1) involves multiple criteria which impose constraints on both the spatial and attribute characteristics of the aggregation outputs, and 2) requires human intelligence to evaluate the trade-off relationships among multiple criteria and determine the 'most appropriate' aggregation solutions. To facilitate the execution of the heuristic aggregation procedure, we also developed a set of visual analytical tools (Figure 19). The toolset has four major components and each of them is used to support one or more step(s) involving human intelligence in the proposed aggregation procedure (Figure 20). The toolset consists of: 1) a scatter plot by which data users can set the threshold(s) for the acceptable error level(s) (aggregation step 1); 2) a parallel plot which presents all aggregation candidate(s) for one seed (the searching result in step 2) and measures of selected aggregation criteria so that users can execute step 3- evaluating the trade-offs between the criteria; 3) a console panel which supports users to work through all aggregation seeds and repeat step 2 and step 3 to determine the "best" aggregation option for every seed; and 4) a set of map windows to present the spatial information to facilitate

the trade-off evaluation. All these tools are linked together so that the information on display is updated synchronically.

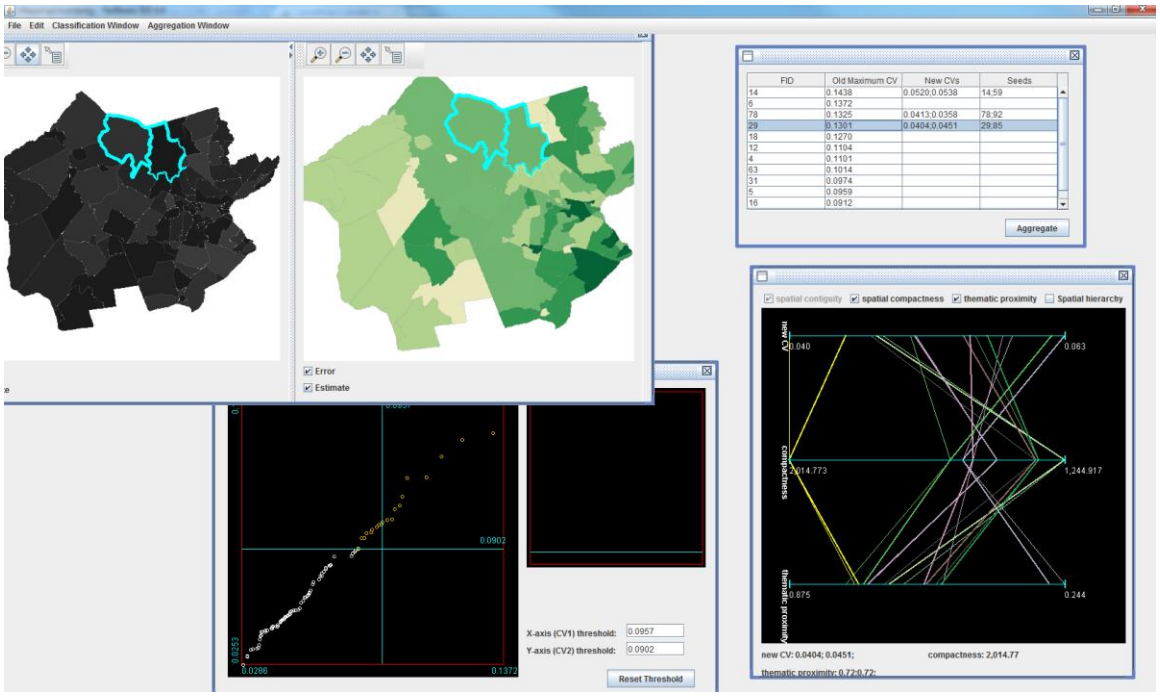

Figure 19 Screenshot of the visual analytical toolset supporting the proposed spatial aggregation procedure (upper left: map windows; upper right: console listing all seeds and candidates; lower left: scatter plot for identifying aggregation seeds; and lower right: parallel plot supporting the selection of best option)

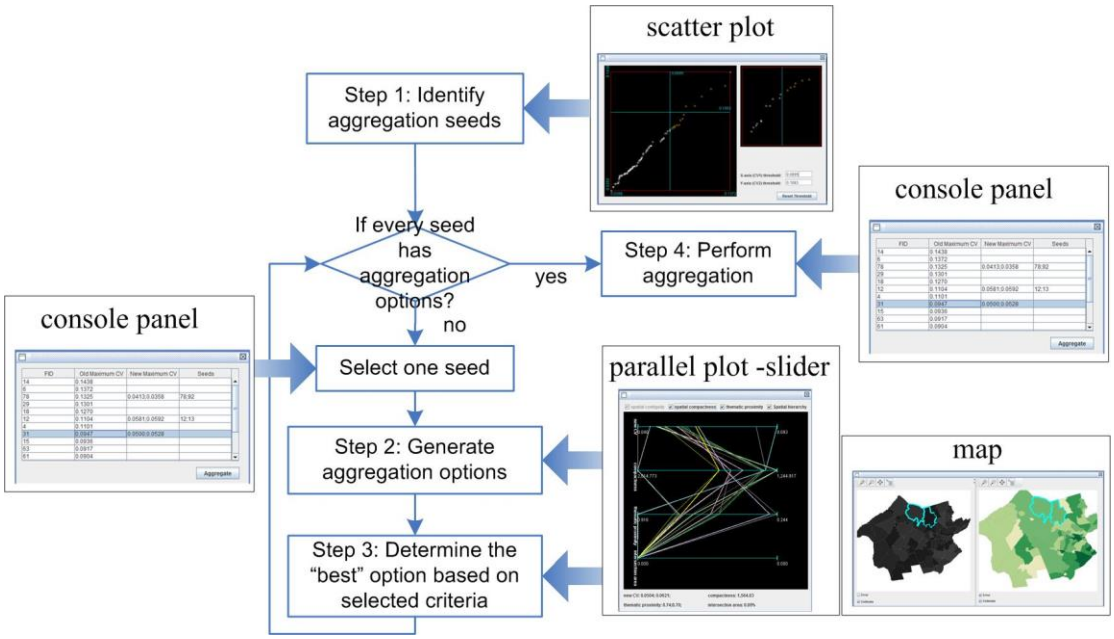

Figure 20 Workflow of executing aggregation procedure and the corresponding supporting tools

4.4.1 Scatterplot View

For users to choose the CV threshold(s), we designed an interactive scatterplot display (Figure 21). If the aggregation is performed on only one variable, all CV values are sorted and display along the horizontal axis. The vertical axis is added to show the CV values of the second variable and their relationships with the first variable if two variables are involved in the aggregation. A zoom-in window is provided to magnify the area with clustered points. The vertical blue line marks the threshold of acceptable error level for the first variable on the horizontal axis, and the horizontal blue line marks the threshold of acceptable error level for the second variable on the vertical axis. By dragging the two threshold lines with the pointer device, users select the lowest CV values that they can accept. After setting the CV thresholds, units on the right of the vertical threshold line are identified as the seeds for aggregation when the aggregation is

taken only on one variable. When two variables are involved, all units found in the upper left, upper right and lower right quadrants (colored in orange) are the seeds for aggregation, because these units have at least one CV value larger than the corresponding threshold. Alternatively, user could just type in the threshold value(s). The geometry id and CV information of the seeds will be updated in the console panel.

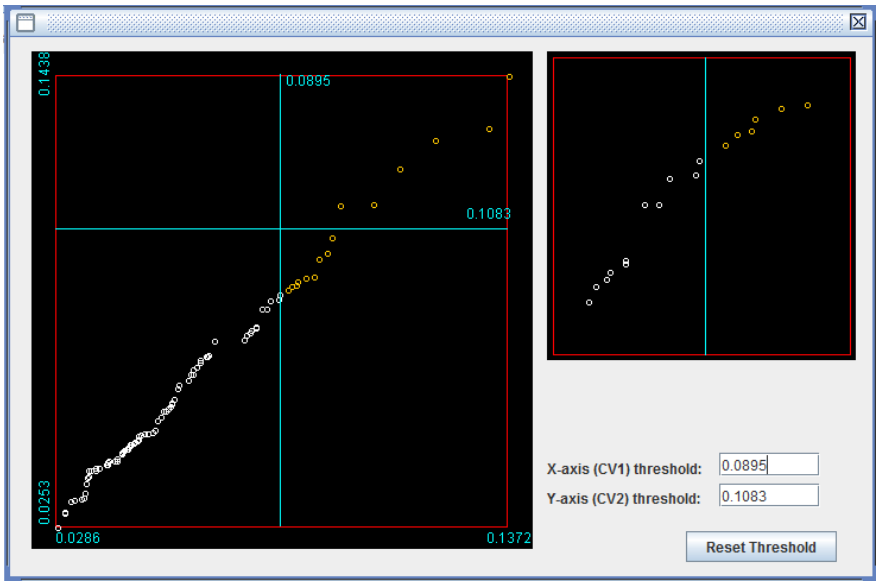

Figure 21 Screenshot of the scatter plot view

4.4.2 Console Panel

A component of the toolset is the console panel, which consists of a table, a series of functions that enable interactive manipulations on the table, and a command button to perform aggregation (Figure 22). The table provides an overview of information about the seeds and associated "most preferable" aggregation options. As soon as users determine the CV threshold(s), seeds for aggregation are identified and displayed on the

table. Seeds are sorted in descending order according to the CV values of all involved variables. Regardless of which variable, seed with the larger CV values are listed first. In the table, each row represents one seed and its related information. From the left to the right, the columns are: 1) the feature id of the seed, 2) the CV of the seed or the maximum CV of the seed if the aggregation involves two variables, 3) the new CV(s) of the new unit for all involved variables, and 4) the feature id(s) of units in the "best" option that will be aggregated with the seed.

The system automatically starts a new search of aggregation candidates/options and calculates optional criteria as soon as the user selects a row (seed) in the table. Note that the user can determine the optional criteria they want to consider before the entire aggregation procedure starts. Meanwhile, location of that seed will be highlighted on the maps. As soon as the user determines the "best" option for the current seed according to the trade-off relationship in the parallel plot (to be discussed below), information including the feature id(s) of units to be aggregated and the CV of the potential new zone will be automatically updated in the third and fourth columns in the table. After the user finishes evaluating various options for all seeds and decides to commit the aggregation, candidate units will be merged to their corresponding seeds to complete the aggregation procedure.

Since selecting the best aggregation options incorporates human intelligence, the user may not be satisfied with the initial choice of the "best" option for a seed. To reverse the possibly "mistake", the console panel supports a roll-back operation such that when the user re-clicks a seed, the records about the "best" option for that seed, if exists, will

be erased from the table. Another round of searching and selecting candidates will start again.

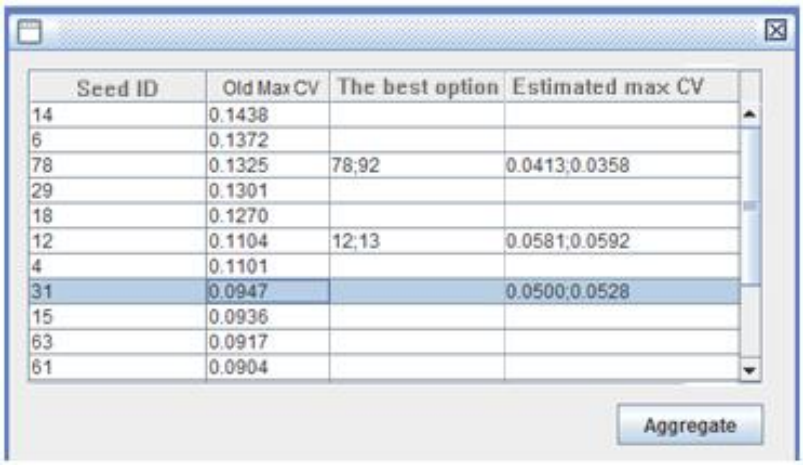

Figure 22 Screenshot of the console panel

4.4.3 Parallel Plot View

Parallel plots have great potential to reveal underlying multivariate relationships. We develop a parallel plot combining slider and filtering techniques to drive the complex evaluation on the trade-offs between multiple criteria in the process of choosing the "most preferable" option (Figure 23). Upon selecting a seed in the console panel, the user actually triggers the process of searching candidates for the seed and computing measures for all criteria for every candidate. The parallel plot is updated accordingly to display the values of those criteria. Except the first axis which represents the derived CV of the merged zone, the other axes are associated with the optional criteria and each axis represents one criterion. Users select optional criteria to be considered in the evaluation by checking the boxes on the top of the plot. Number of axes displayed on the parallel

plot is adjustable according to the number of selected criteria. Values of criteria arrange from left to right, reflecting desirable to undesirable conditions.

To indicate the trade-off relationships between different criteria, values for different criteria associated with the same candidate are connected by line segments (Figure 23 left). A polyline represents an aggregation candidate that scores on different criteria. Colors of the polylines are used to distinguish different candidates. If aggregation involves only one variable, every candidate corresponds to one polyline. When two variables are involved, every candidate should have two sets of criteria values. We use polylines of the same color but in different widths to distinguish the two sets of criteria values associated with the same candidate.

To alleviate crowdedness in parallel plots with a large number of candidates or lines, we add a slider with two cursors delimiting the minimum and maximum values of every axis on the plot. By moving the cursors on the axis, users can control the value range acceptable for that criterion. Candidates with criterion values outside the range will be filtered out and colored in dark grey (Figure 23 right). Thus, users can narrow their choices to the remaining candidates.

Data brushing and linking techniques are also implemented in this view. Individual candidates can be highlighted on the plot, and corresponding criteria values are shown underneath the plot to enable accurate comparisons (Figure 23). Maps are updated highlighting the candidate areal units, so that users can consider the candidate locations and the seed. After the user finishes evaluating various options and selects the

preferable candidate, related information of the candidate (including geometry ids) will be recorded in the console panel.

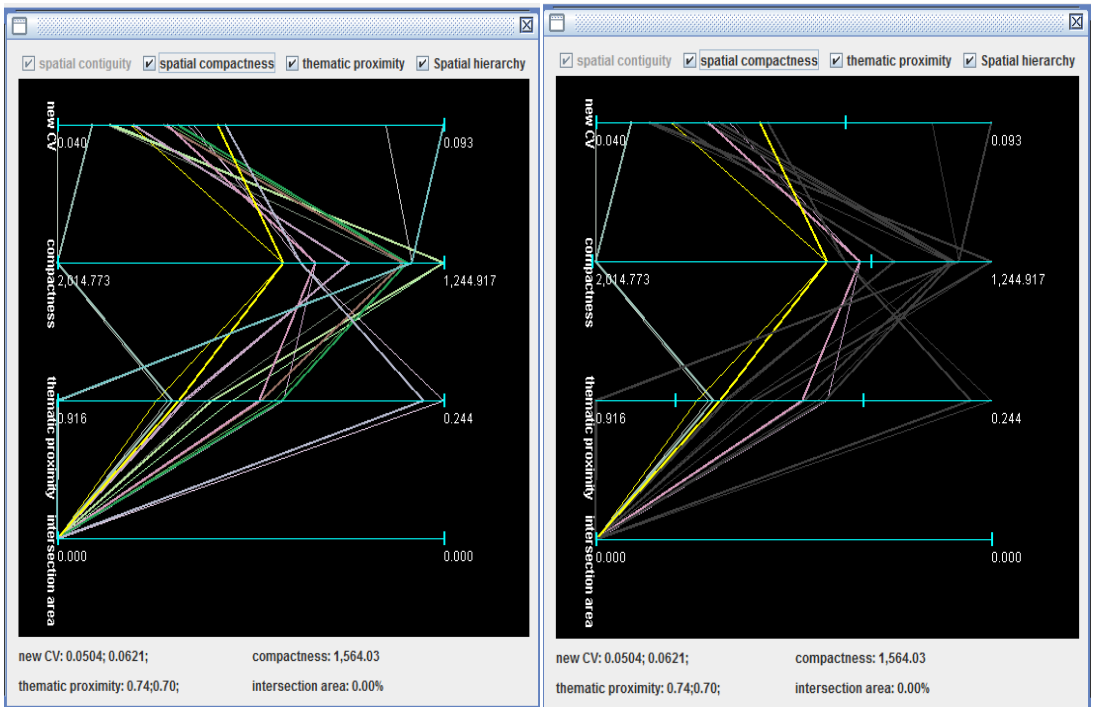

Figure 23 Screenshots for parallet plot view: (a) parallel plot presenting all candidates; (b) cursors were used to narrow the ranges of the first three axes such that some candidates were grayed out

4.4.4 Map View

Maps are used to show the locations of areal units and spatial patterns of attributes in the heuristic aggregation process (Figure 24). Maps are generated for both the estimates and errors of involved variables, and are housed inside a scalable frame. Two windows are in the frame, and each window corresponds to one variable. Each window includes two map layers, the estimate (bottom layer) and the error (top layer). Users can set the visibility of every layer through the check boxes. To see the bottom layer, the top

layer has to be set as invisible. Moreover, users are allowed to minimize any window (either the one on the left or right) in order to focus on the other one. Choropleth map is used to display the spatial pattern of CV. Presenting estimates using a choropleth map may not be meaningful due to large error in estimates, as estimates in different classes may not be significantly different. An unclassified map with no clear class break values might be more appropriate to avoid portraying the impression that estimates in different classes are statistically different (Tobler 1973). Therefore, estimates are presented using an unclassified map. Since maps are linked to all the other views in the toolset, areal unit(s) will be synchronously highlighted on maps in response to any selection made in other views.

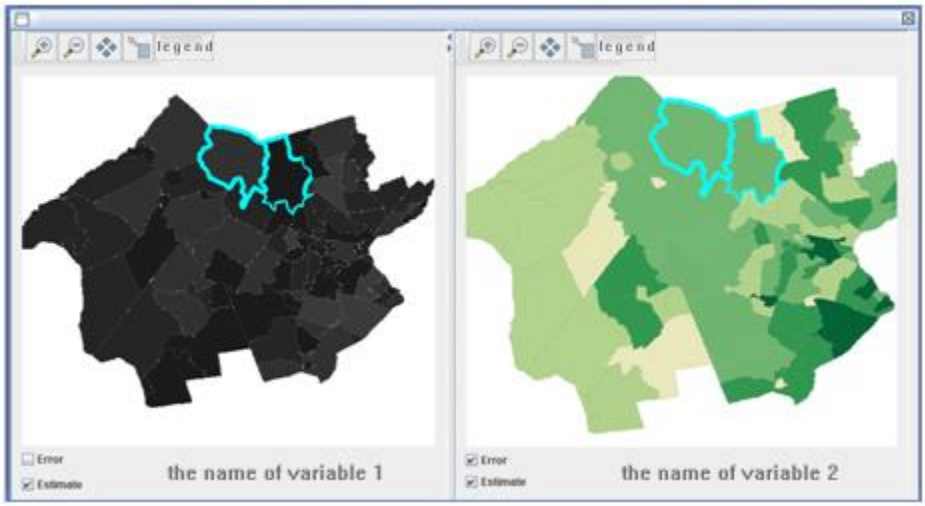

Figure 24 Screenshot of the map view (in the map view example, left frame displays the unclassified map of the estimate of the variable 1, and the right frame displays the choropleth map of the error of the variable 2)

In summary, a heuristic and flexible aggregation procedure with user inputs in determining the best merging configuration based on the constraints from multiple

criteria provides a possible solution to reduce error and produce data with more usable estimates. I described the procedure of performing such aggregation on one or two variables and the supporting visual analytical tools in this chapter. Data with large errors in estimates will be used to test the proposed aggregation procedure in the next chapter.

CHAPTER FIVE DEMONSTRATING THE PROPOSED MAPPING AND AGGREGATION METHODS

As error should be considered in choropleth mapping, Chapter 3 proposed a class separability measure and two map classification methods: maximizing class separability and considering separability and other criteria through a heuristic process. In this chapter, I will demonstrate the proposed mapping methods, and evaluate their effectiveness and deficiencies. On the other hand, aggregation was proposed as a possible way to improve data quality by creating new estimates with smaller errors. I will perform the proposed aggregation procedure (see Chapter 4), and then compare the resultant data with the original ones to detect the magnitudes of changes in estimates and errors through aggregation. Theoretically, all proposed approaches should be applicable to many datasets which provide error information (measured by MOE, SE or CV) for each estimate. Considering the popularity and importance, data from ACS and the national mortality data were chosen for the demonstrations.

5.1 Data to Be Used in the Demonstrations

5.1.1 The American Community Survey (ACS) Data

The ACS is a continuous measurement program administered by the Census Bureau as the substitute of the long form for the decennial censuses on 2000 and earlier. Unlike the decennial census which is conducted every ten years, the ACS questionnaire is mailed to approximately 250,000 household addresses across selected counties in the

U.S. every month. Over the year, the total sample size is about 3 million addresses, which is about 2.3 percent of the total 129.5 million household addresses in the U.S. in 2005. While this sample is relative large as compared to major national surveys, it is still small as compared to the approximately 18 million addresses (or one-sixth of total household addresses) which received the census long form in 2000. The uncertainty of estimates can thus be an issue in the analysis using ACS data (MacDonald 2006). Simple absolute differences between estimates may not be regarded as real differences because estimates are just means from sample observations. Therefore, when analyzing the spatial relationships or spatial distributions of ACS estimates, one should consider to what extent the differences among estimates are real and observations assigned to different map classes have different estimates to ensure the authenticity of derived information presented by the choropleth maps.

Due to the sampling scheme specific to ACS, Census Bureau categorizes ACS products by the length of period from which estimates are derived. The one-year estimates have been released to the public since 2006, and the first set of three-year estimates was released in 2008 while the first set of five-year estimates in 2010. Estimates are available for administrative and statistical units from census block group up to the nation. The Census Bureau provides one-year estimates only for counties or larger administrative units with at least 65,000 people. In the three-year data products, the smallest reporting areas, census tract, need to meet the minimum threshold of 20,000 people. The geographical units of the five-year estimates are as small as census block groups. Therefore, choosing which data product is partly dependent upon the census

statistical or administrative units to be mapped or analyzed. The use of ACS data is likely to increase tremendously as they replace the long form, providing the only source of socioeconomic data of the population in the U.S. Therefore, we choose ACS as one of the data sources for the demonstrations below. In ACS, MOEs are provided in a column following each column of estimates of a variable in every available table.

5.1.2 National Mortality Data

The U.S. mortality data are collected and maintained by the National Center for Health Statistics (NCHS) starting as early as 1961 (The United Nations 2004). The data are one of the few sources that describe the proportion of deaths and the cause of death in a population over a period of time for small geographic areas. Public users may access the mortality datasets through the Surveillance, Epidemiology, and End Results (SEER) Data Management System[3].

The mortality data cover the entire U.S. population. Deaths attributable to various causes are gathered by states. The sample size of mortality data is equivalent to the response rate. The sample size/response rate varies by geographical regions and mortality topics (such as infant deaths or death in the other age groups). Till 2007, over 2.4 million deaths have been reported in the U.S.

In the SEER system, mortality data include the data for the period from 1969 to 2011. Health Service Areas (HSAs) are one of the basic units used to report estimates of mortality data. HSAs were originally defined by NCHS and each consists of a single county or a cluster of contiguous counties, which is "relatively self-contained with

[3] http://seer.cancer.gov/seerdms/

respect to the provision of routine hospital care" (Makuc et al. 2001; Town et al. 2007). Because error occurs during data collection, standard errors (SEs) are provided with estimates in the database. Mortality data have been widely used in a variety of research activities, such as determining life expectancy and disclosing mortality trends, etc. An atlas was compiled to present the mortality and related information in the U.S. (Pickle et al. 1996). Due to the easy access and frequent use, the mortality data were chosen as another data for our demonstration below.

5.2 Evaluating Classification Methods

Four variables selected from ACS and SEER databases are used to demonstrate the effectiveness of the proposed mapping methods. Selected descriptive statistics of these four variables are reported in Table 1. To compare the reliability of estimates with a wide range of values, the CV instead of SE/MOE should be used, as CV is the SE divided by the estimate, removing the scale effect of the estimates. Note that since some CV values for the four datasets are very small, we multiplied the CV values by 100 percent. Among the four datasets, the U.S. county-level data of median household income have the largest CV, implying that they are the least reliable. By contrast, the VA county-level data of median household income and the U.S. mortality rate of white have the lowest CV values, implying that the estimates of both variables are relatively reliable.

Table 2 Summary statistics of the four variables: N = number of observations, MOE = margin of error, SE = standard error, CV = coefficient of variation (in %), Min = minimum, Max = maximum, STD = standard deviation (0 is rounded from some small non-zero value).

Data Sets	N	Mean	Average MOE	Average CV	Min CV	Max CV	STD CV
(1) Median household income, Iowa counties, 2006-2010 ACS	99	46,437	2,571	343	94	954	148
(2) Median household income, Virginia counties 2009-2011 ACS	134	55,168	3,441	4	1	10	2
(3) Median household income, U.S. counties, 2007-2011 ACS	3,109	45,110	2,893	419	26	4,400	305
(4) Mortality rate of white (rate per 100,000 people), U.S., 1969-2004, SEER	798	988	12	1	0	3	0

5.2.1 Evaluating Conventional Classification Methods Using Separability Measure

As proposed in Chapter 3.3.1, the class separability concept can be used to evaluate how effective conventional classification methods can determine classes such that estimates in different classes are really different statistically. Three popular classification methods, Jenks natural breaks, equal interval, and quantile, are selected to be evaluated. Map classes were determined using the three methods for three different data sets (the county-level estimates of median household income for Iowa, Virginia and the U.S). Figure 25 shows the maps and legends made using the Iowa county-level data. In the natural-breaks map, the class break between the fifth and sixth classes has the

highest separability level. The estimates on the two side of the break are statistically different at 66 percent or higher confidence level (CL). However, the lowest separability is associated with the class break between the third and fourth classes and only about 8 percent. Compared to the natural breaks methods, only one class break generated by the equal interval method has the CL exceeding 30 percent, and the CLs of all class breaks generated by quantile are lower than 10 percent. Table 2 reports the separability levels/CLs and the averaged values by classification methods and datasets. While the results clearly show that the quantile method performs the most poorly, results are inconsistent whether natural breaks or equal interval is better. Also, these classification methods can produce more separable classes when data (the Virginia county-level data) contain relatively lower errors in estimates.

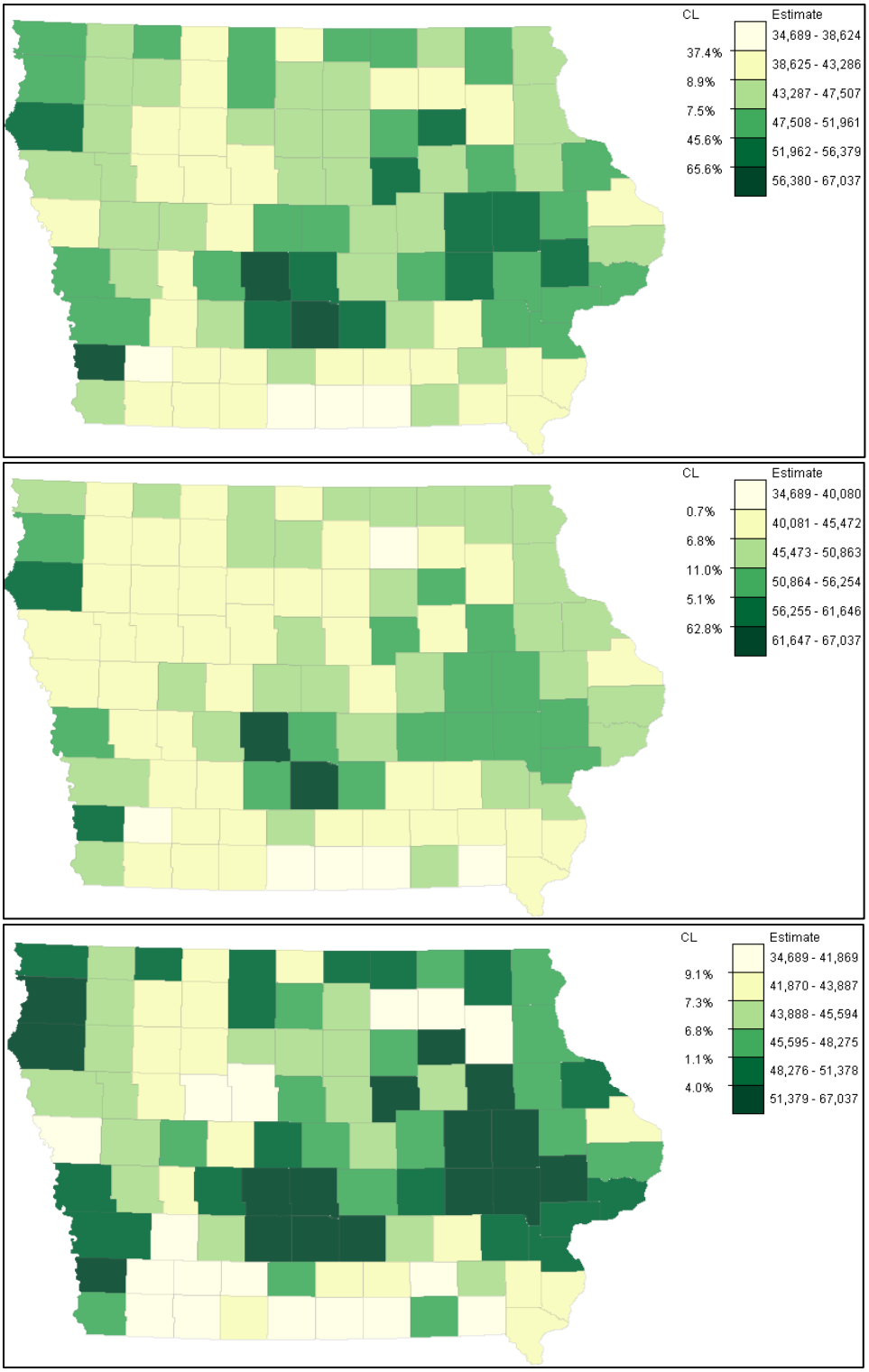

Figure 25 Choropleth maps of Iowa county-level estimates of median household income made by traditional classification methods (Top: Jenks natural breaks; Middle: Equal interval; and Bottom: Quantile).

Table 3 Confidence levels associated with class breaks for the four classification methods, using the three American Community Survey (ACS) datasets. (CS: class separability; NB: Jenks natural Breaks; EI: equal interval; Q: quantile; percentages are rounded, and therefore 0% are actually less than one)

	Median Household Income of Iowa Counties, 2006-2010 ACS: Approximated Confidence Levels (in %)				*Median Household Income of Virginia Counties, 2009-2011 ACS: Approximated Confidence Levels (in %)*				*Median Household Income of U.S. Counties, 2007-2011 ACS: Approximated Confidence Levels (in %)*			
Breaks between Classes	*CS*	*NB*	*EI*	*Q*	*CS*	*NB*	*EI*	*Q*	*CS*	*NB*	*EI*	*Q*
1-2	47	37	1	9	92	30	20	32	62	1	1	0
2-3	46	9	7	7	88	32	23	24	70	1	2	0
3-4	66	8	11	7	100	30	92	17	69	3	12	0
4-5	63	46	5	1	100	92	100	8	97	5	36	2
5-6	99	66	63	4								
Mean	64	33	17	6	95	46	59	20	75	2	12	1

The separability concept and the associated legend design provide statistical meaning to class breaks such that map readers have a clearer idea about the extent that enumeration units in different map classes have statistically different values. From a map interpretation perspective, this information is important when map readers intend to compare values across units to draw conclusions regarding the potential spatial patterns.

We can also evaluate the performance of different classification methods by measuring their robustness. We build on the robustness concept suggested by Xiao et al. (2007), but remove the requirement to specify a threshold probability. We define robustness of each areal unit as the probability that it is classified correctly given the error distribution for its estimate. For an estimate in a given class, its distribution (PDF)

stretches across to other classes, as shown in Figure 2 (see Chapter 2.2.1). Area under the PDF bounded by the class interval indicates the probability that the estimate is assigned to a correct class (i.e. robustness). This probability should be relatively high for a robust classification method and it can be determined for each estimate. To evaluate the classification performance, we take the average robustness of all estimates.

The robustness levels of the three classification methods using the three datasets are reported in Table 4. For comparison, the averaged separability levels (CLs) of all classification methods are also included. Although these results evaluated by robustness are not completely consistent with the evaluation using separability (Table 4), the quantile method fares the worst according to both evaluation criteria. The natural breaks and equal interval methods are quite similar in their performances, but the equal interval method seems to have a small edge over the natural breaks method most of the time. To some degree, these results are consistent with the evaluation performed by Xiao et al. (2007) in which the quantile method had the worst performance in their state-level data.

Table 4 Comparing the performance of the four classification methods using the three American Community Survey datasets based upon the averaged robustness measure in Xiao et al. (2007) and the proposed averaged confidence level in percent. (CS: class separability; NB: Jenks natural breaks; EI: equal interval; Q: quantile)

Data Sets	*Performance Measures*	*CS*	*NB*	*EI*	*Q*
Median household income, Iowa counties, 2006-2010 ACS	Robustness	0.89	0.60	0.64	0.51
	Separability	64	33	17	6
Median household income, Virginia counties 2009-2011 ACS	Robustness	0.97	0.83	0.89	0.8
	Separability	95.0	46.2	58.8	20.0

Median household income, U.S. counties, 2007-2011 ACS	Robustness	0.98	0.8	0.88	0.72
	Separability	75	2	12	1

5.2.2 Evaluating Maps Created by Maximizing Class Separability

In addition to evaluating classification methods, we can also determine classes based on the separability levels between estimates to create the most separable classes. In Chapter 3.3.2, I already presented the process of determining highly separable classes supported by the proposed set of visual analytical tools (the bar plot and slider bar). Here, real world datasets are used to show how the process and the toolset operate.

Figure 26, Figure 27, Figure 28 and Figure 29 present the process of determining class breaks by maximizing separability using the Iowa county-level data. The process involves evaluating the trade-off between the lowest acceptable separability level and the largest number of class to be created. Since high separability/CL is more desirable, we start the mapping process with a relative CL high value as the lowest separability level that can be accepted. By moving the cursor on the slider bar to 80% as the lowest acceptable level (i.e., estimates separated by the class break are statistically different at 80% confidence level), the data produced only two classes, which are probably not enough for the data with 99 observations (Figure 26). Also, the resultant classification is highly unbalanced with one observation in the higher class and all the other observations in another class. To obtain more map classes, we have to lower the threshold of lowest acceptable separability by moving the slider bar further to the left. If the lowest separability level is down to 60%, then four classes are derived (Figure 27). Although the

current number of classes is much more reasonable than the last trial, the classification is still highly unbalanced. When the lowest separability level is decreased to 45%, six classes are created, and seven classes emerge if decreasing the level to 40% (Figure 28 and Figure 29). Compared to six classes, the legend for seven classes is more difficult to read. When increasing the number of classes from five to six, the resultant classification is more balanced than before. A group of observations was separated out from the largest group as a new class. However, when increasing the number of classes from six to seven, the unbalancing situation was not improved further. Considering the ease of recognizing classes, the imbalance of observations among classes and separability levels, we can conclude that the six-class scheme is better than the five or seven-class schemes.

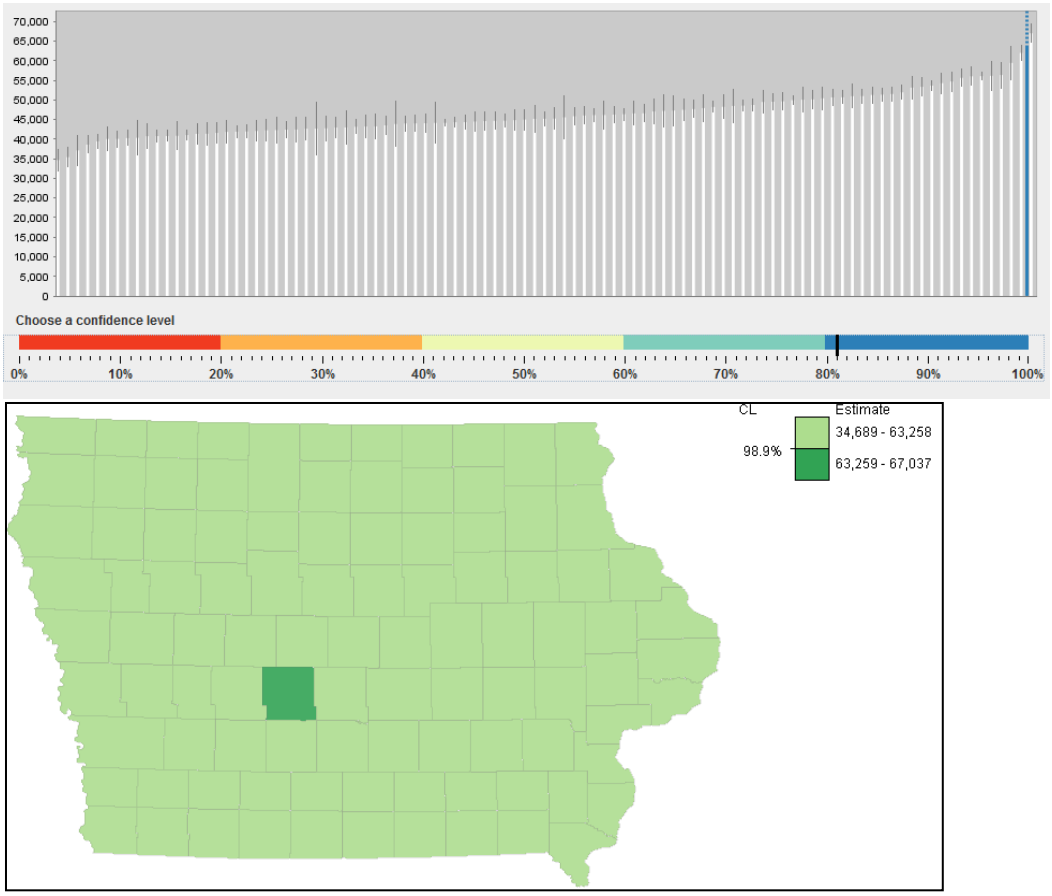

Figure 26 Map generated by accepting the minimum separability as 80%

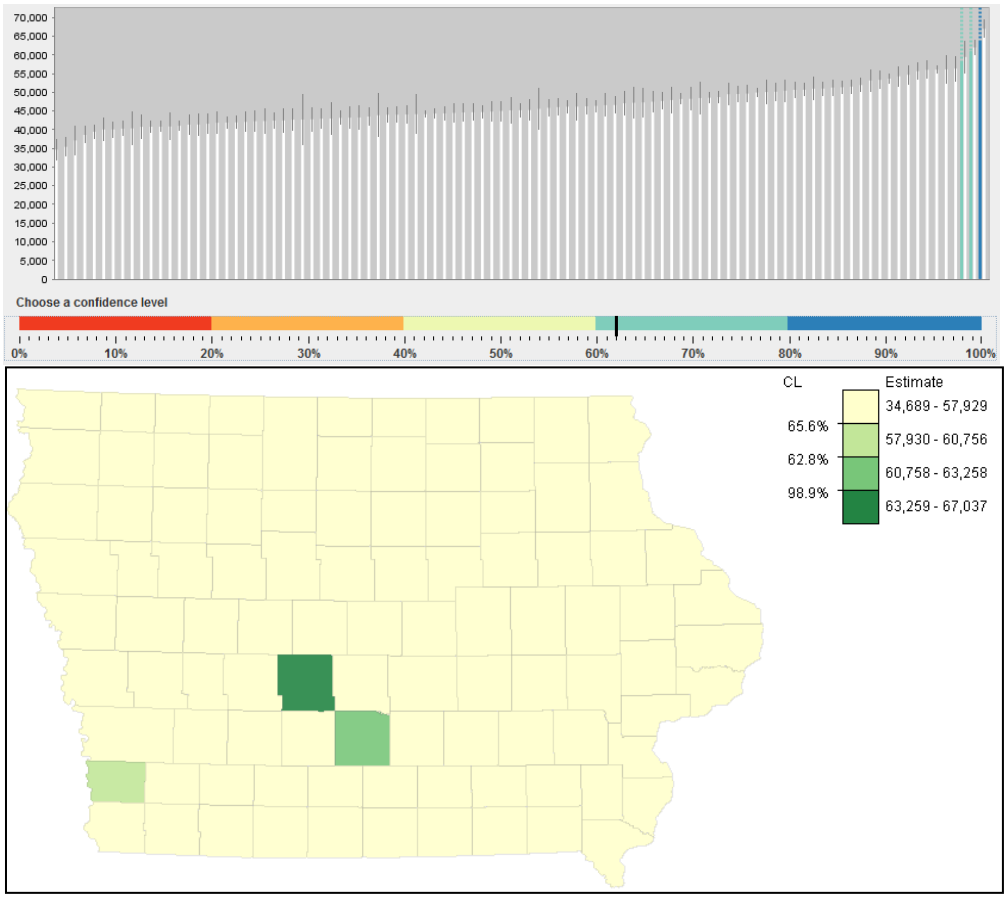

Figure 27 Map generated by accepting the minimum separability as 60%

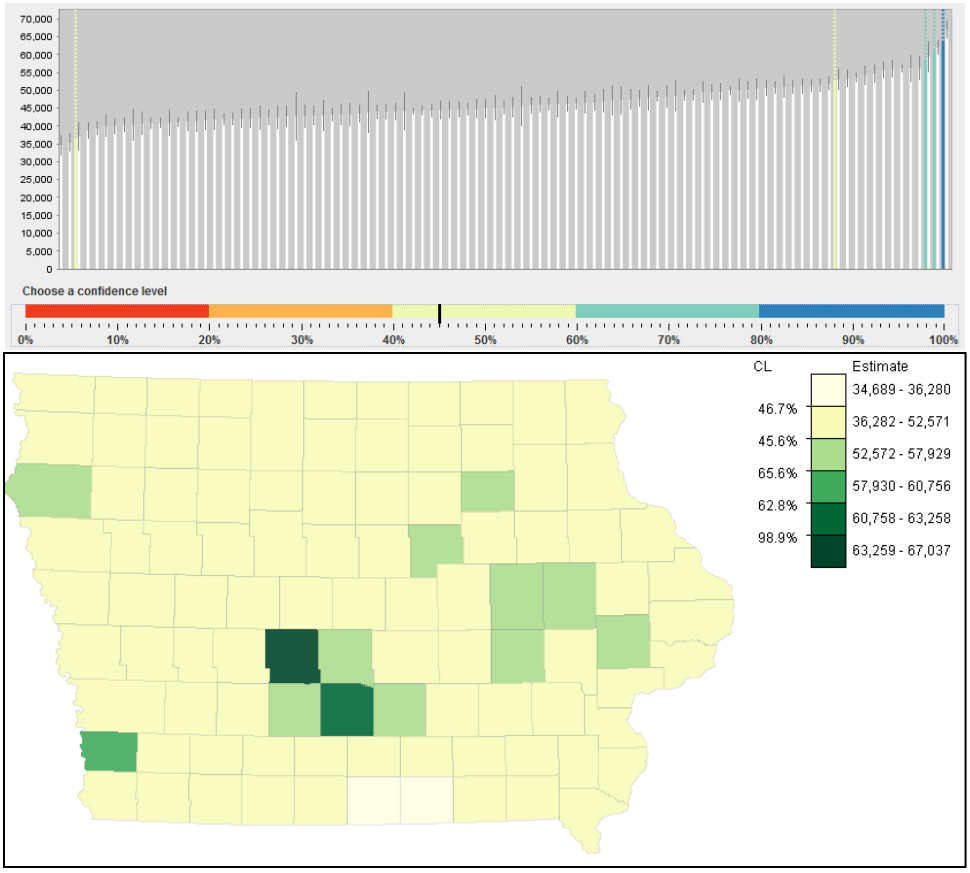

Figure 28 Map generated by accepting the minimum separability as 45%

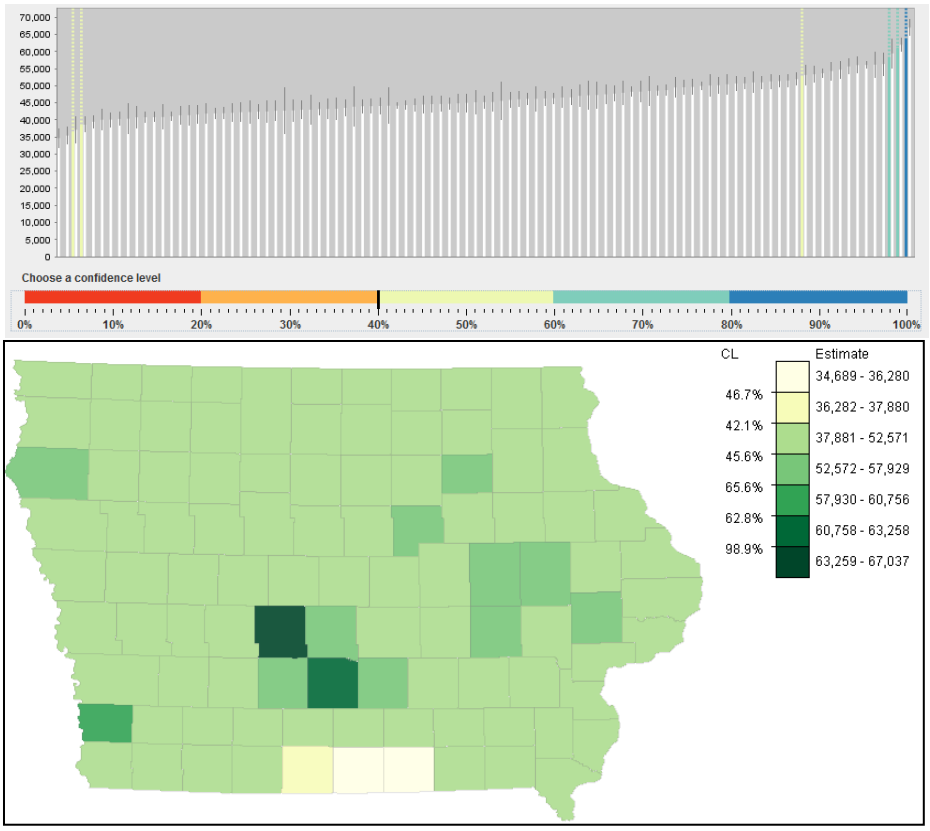

Figure 29 Map generated by accepting the minimum separability as 40%

Therefore, in the Iowa example, we chose six classes, assuming that the user was willing to accept a minimum separability level of approximately 45 percent (equal to the level between the second and third classes from the top on the legend, Figure 4 lower). The highest separability level approximates 99 percent, creating a class break between the fifth and sixth classes. The second highest separability level is around 66 percent creating another class break between the third and fourth classes. Subsequent class breaks have successively lower separability levels. By interpreting the legend, one may say that estimates divided by the last class break value are approximately 99 percent statistically different, while values separated by the first break are only 47 percent statistically

different. Because of the relatively high separability levels of the break between the two highest classes, we have more confidence to say that counties with the highest values in the south-central part of the state are different from their surrounding counties and the rest of the state assigned to other classes. For those counties with medium values scattered across the state, the chances that they are different from the counties in other classes are much lower, only around 50 percent.

Besides the Iowa data used above, the Virginia and U.S. county-level data of median household incomes are also examined through choropleth maps (Figure 30 and Figure 31) using the class separability classification method. In the case of the U.S. county-level data, even if we are willing to lower the separability to about 10 percent, only eight class breaks could be determined with the highest separability level at only about 24 percent. With such low separability levels for most of the observations, one should consider if assigning observations to classes is meaningful at all, given that those values are not different in statistical sense. In contrast, the five-class map of Virginia county-level data has higher separability level associated to every class break. The lowest separability of the break between the second and third class is as good as 88 percent. Meanwhile, selected statistics of the two datasets (Table 2) show that the CV values of the Virginia data range from 1 percent to 10 percent, but the CV values of the national data range from 26 percent to 4,400. The U.S. county-level data have much larger errors in their estimates than the Virginia data.

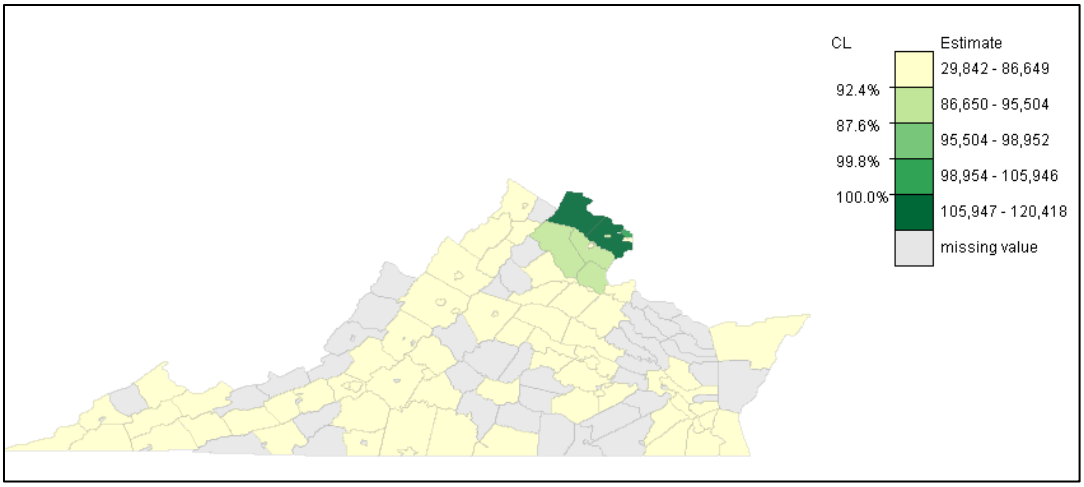

Figure 30 Choropleth map of Virginia county-level estimates of median household income with classes determined by maximizing separability

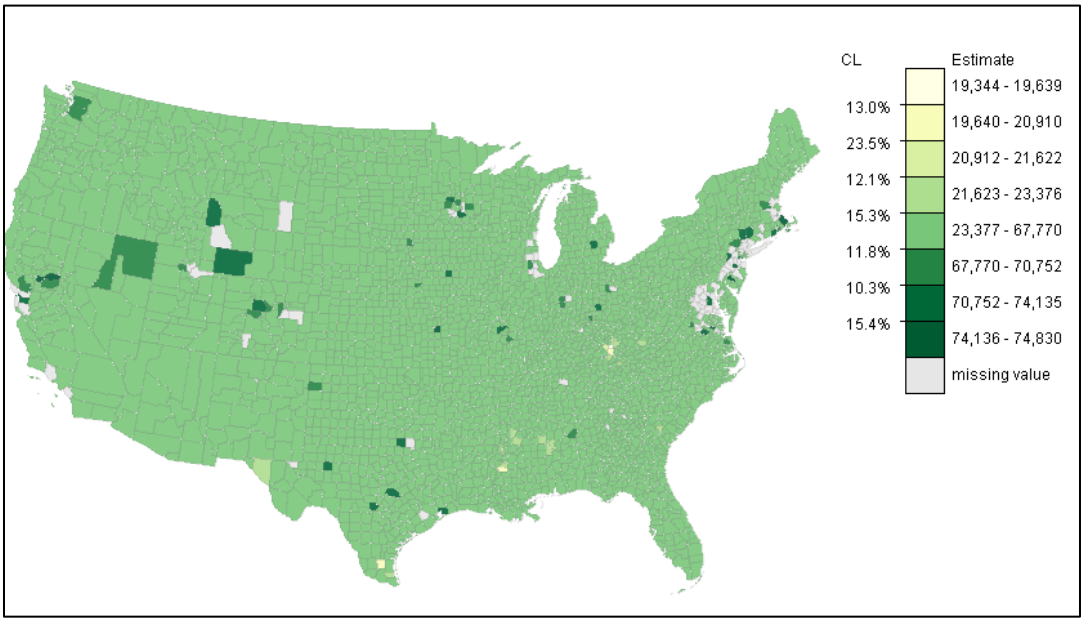

Figure 31 Choropleth map of National county-level estimates of median household income with classes determined by maximizing separability

The results of mapping Virginia and the U.S. country-level data demonstrate that the separability classification method is a data driven method. If estimates have relatively large errors, then separability levels will generally be low. Nevertheless, the class

separability method should be able to identify breaks with the highest possible separability levels, given the number of desired classes. Cartographers have some control over the trade-off between class separability levels and the number of classes. However, the usefulness of such maps needs to be evaluated further involving map users-readers. Moreover, in the case of mapping data with large error, should the separability method or any classification method be used? The unclassed map with no clear class break values may be more appropriate to avoid portraying the impression that values in different classes are different (Tobler 1973).

In addition, we compared the classifications created by maximizing class separability with the ones determined using conventional methods. We used both the proposed separability measure and Xiao's robustness in the comparison (Xiao et al. 2007; Table 3 and Table 4). Apparently, the class separability method performs better than other methods according to both evaluation criteria. This result should not be too surprising as popular classification methods do not take into consideration the errors in estimates, and therefore statistical differences between values are not of concern.

5.2.3 Heuristic Mapping Results

Maps made by maximizing class separability classification may have a potential problem: observations may be unevenly assigned into different classes. Datasets with positively skewed distributions are very common. Observations at the right-hand tail are often statistically different from other observations and they may form singular-unit classes, while all the other observation may fall into the same class. This kind of imbalanced classification results may not be useful to map readers to detect the

underlying spatial patterns. To address this issue, we have proposed a heuristic mapping procedure which involves the inputs of human intelligence in determining class breaks by considering separability and other criteria (e.g., within class variation and numbers of observations among classes, see Chapter 3.3.3). The U.S. HSA-level data including the estimates of all cause-of death (COD) mortality rate of white obtained from SEER database are used to demonstrate the effectiveness of the heuristic mapping method.

The heuristic mapping method is supported by an interactive star plot (see Chapter 3.4.2). Assuming that the mapping purpose is to expose the spatial pattern according to the distribution of estimates, we may start the procedure with Jenks natural breaks classification method. The possible number of classes can be from two to nine for the nation-wide dataset with 768 observations (HSAs). Classification schemes are made using the Jenks natural breaks method and different number of classes. In this demonstration, we consider a total of three criteria: separability, within-class variation and number of class. Therefore, a star plot with three axes is used to present the potential classification schemes, and each scheme is a combination of parameter values according to the three selected criteria Figure 32).

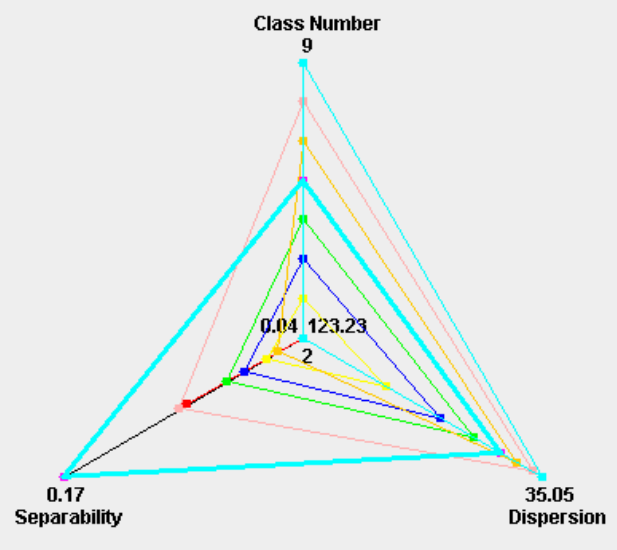

Figure 32 Star plot to support heuristic mapping method

While most schemes have moderate to undesirable values on the separability level (i.e. classification robustness), the robustness for the scheme with six classes is 0.17, which is the highest among all schemes. Although the within-class variation (dispersion) for the six-class scheme is not the best (smallest), it is very close to the smallest dispersion level among all schemes. However, the nine-class scheme has the lowest robustness value and the map with nine classes is not easy for map readers to interpret. Therefore, after evaluating the trade-offs among the three criteria, we selected the classification scheme with six classes. The corresponding map is shown in Figure 33 (upper). To compare the results of the heuristic mapping procedure with the other mapping methods, we also mapped using the class separability (Figure 33 middle) and the 5-class natural breaks methods (Figure 33 lower). The five-class scheme is commonly used in choropleth maps, but the selection of this number is arbitrary without considering any of these mapping criteria explicitly. The separability levels of class breaks are

reported in Table 5 by mapping methods. The averaged separability levels and the values of robustness (Xiao et al 2007) are included in the table.

Table 5 The separability and robustness levels for each class break by classification method (the separability levels are rounded)

Classes	*Natural Breaks*	*Class Separability*	Heuristic
1-2	56	100	44
2-3	8	99	18
3-4	18	95	17
4-5	8	93	23
5-6			64
Mean	22	97	33
Robust	0.95	1.0	1.0

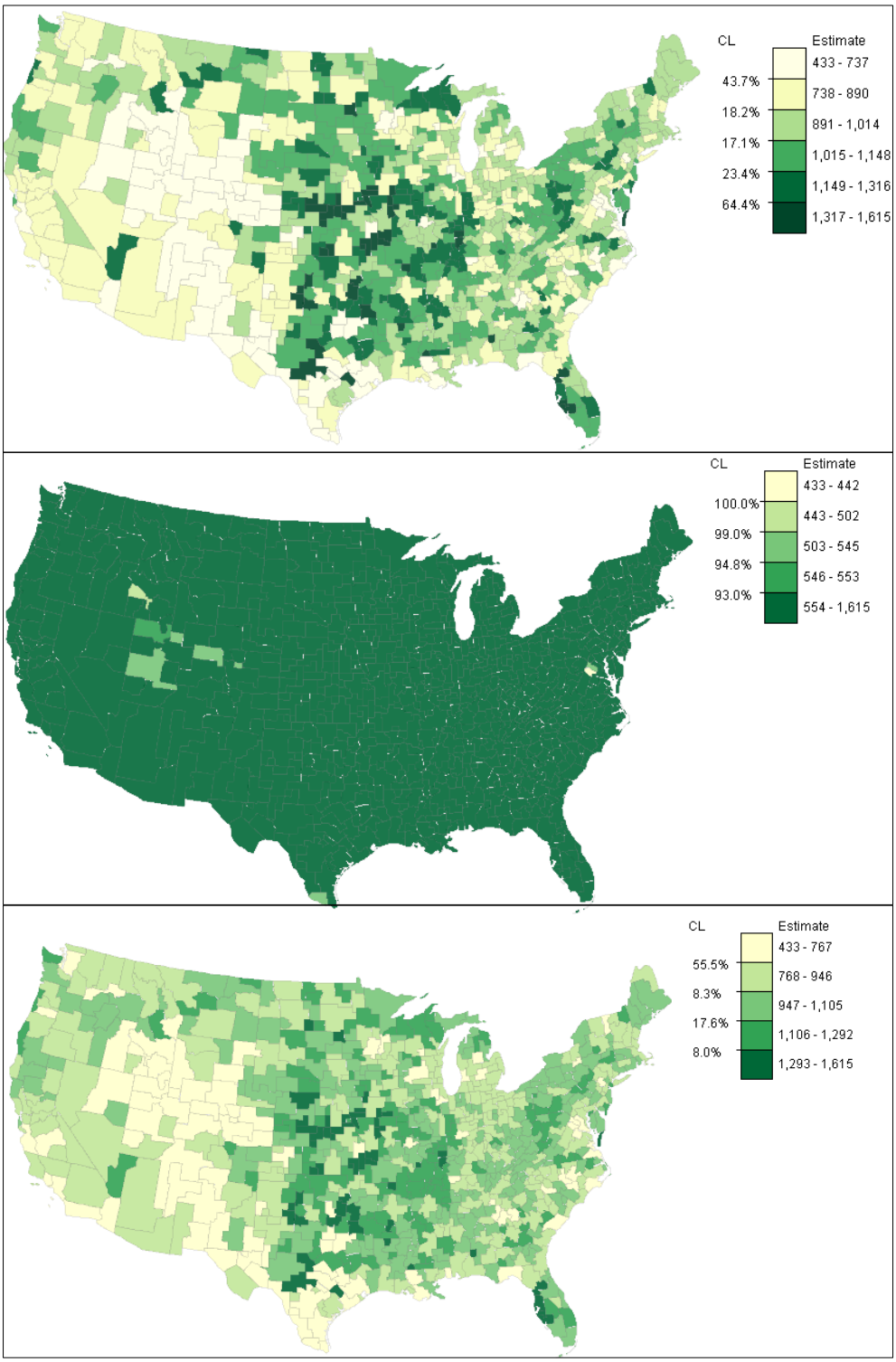

Figure 33 Maps of national mortality rate made by the heuristic method (upper), maximizing separability (middle), and traditional Jenks natural breaks (lower)

As expected, the map created by maximizing separability has higher separability levels than other maps. Due to the relatively small errors in estimates (Table 2), the lowest separability level of the class break between the fourth and fifth classes is still higher than 90 percent. However, most of the enumeration units were assigned into the highest classes. This classification is useful to recognize those observations which are significantly different from other observations, but is not very efficient to represent the spatial distribution of the phenomenon.

In the map created by the heuristic method, the class break between the fifth and sixth classes has the highest separability level, 64.4 percent. The second highest separability level is 43.7 percent, associated with the break between the first and second classes. The separability levels of the breaks creating the middle classes are around 20 percent, and the break between the third and fourth classes has the lowest separability level, 17.1 percent. Compared to the map made using the standard natural breaks method with five classes, the heuristic method improved the average separability from 22.4 percent to 33.4 percent.

In cases of studying mortality, the classes with the high estimates are usually the group that is the most concerned about by map readers, as these groups reflect the population at risk. However, in the five-class map made by the standard natural breaks method, the separability level between the two highest classes is only about 8 percent. It is to say that the estimates separated by this class break are statistically different only at 8 percent confidence level or lower. In other words, many estimates assigned to the fourth and fifth classes are not statistically different. Similarly, the confidence level between the

fourth and third classes is only 17.6 percent. In order to highlight areas with high mortality rates at a higher confidence level, the separability levels of breaks between classes with high estimates have to be improved. In the map created by the heuristic procedure, the separability level between the highest classes (fifth and sixth) was improved dramatically to 64.4 percent. The map shows with higher levels of confidence that areas with the highest mortality rates cluster along the middle corridor of the nation between Louisiana and Iowa-Nebraska. Based on the above comparison, we can conclude that the heuristic mapping method created more useful results than the other methods to meet the purpose of revealing spatial patterns with a certain confident level.

5.3 Evaluating the Proposed Aggregation Procedure

The proposed classification methods have been demonstrated to be effective of putting statistically different estimates into different map classes, given data quality are moderate to good. Below, I will show how the proposed heuristic aggregation procedure can generate new but more usable data from estimates with large error. Two ACS variables, the female and male population counts of census tracts in Hunterdon county and Somerset county of New Jersey were selected as the inputs of the aggregation demonstration. Selected descriptive statistics of the two variables are reported in Table 6. The two variables have similar reliability levels. Both of them have some observations with CV values larger than 10 percent, which means the SEs of these estimates were larger than one tenth of the estimates. If we consider 9.5 percent (set the value slightly lower than 10) as the highest acceptable threshold for error, then aggregation will be performed on all observations with CV larger than 9.5. For male and female population

counts, 12 areal units have at least one estimate with CV larger than the threshold. The 12 areal units are selected as the aggregation seeds (Figure 34).

Table 6 Summary statistics of the two data sets used in aggregation experiments: N = number of observations, MOE = margin of error, CV = coefficient of variation (in %), Min = minimum, Max = maximum, STD = standard deviation.

Data Sets	*N*	*Mean*	*Average MOE*	*Average CV*	*Min CV*	*Max CV*	*STD CV*
Female population, census tracts in Hunterdon county and Somerset county, New Jersey, 2007-2011 ACS	94	2,436	217	5.9	2.9	13.7	2.3
Male population, census tracts in Hunterdon county and Somerset county, New Jersey, 2007-2011 ACS	94	2,348	221	6.2	2.5	14.4	2.4

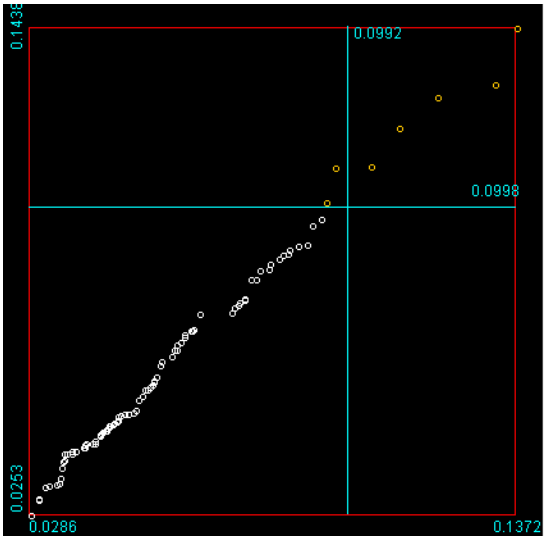

Figure 34 Scatter plot supporting data users to select aggregation seed (the orange points are the aggregation seeds)

Spatial compactness and thematic similarity are chosen as the optional criteria to guide the selection of candidates to be merged with aggregation seeds. When evaluating the trade-offs between the optional criteria and the reduction in error after aggregating estimates, our general strategy is to select candidates which have desirable or moderate values in all criteria for both input variables. Graphically, the best candidate should have a polyline closest to the left side of the axes (Figure 35). Supported by the suite of visual analytical tools, we iterate all aggregation seeds and select the candidate to be merged with each of the seeds. The aggregated male and female counts were derived and the associated MOEs were estimated for the new areal units (Equation 15, see Chapter 4.3).

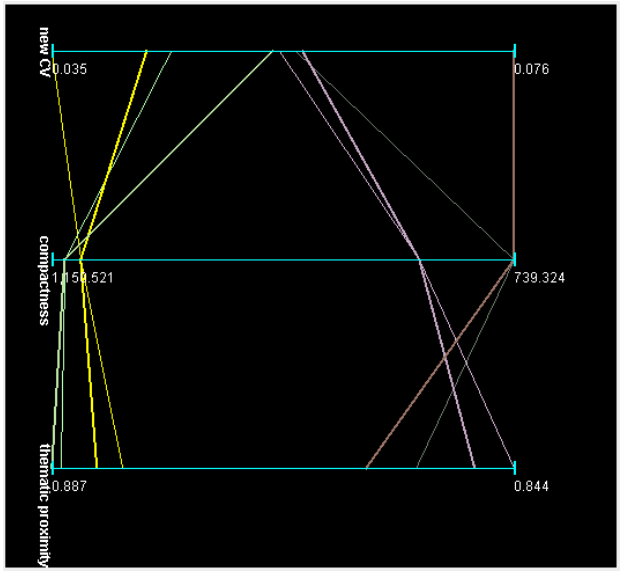

Figure 35 An example of parallel plot used for selecting the "best" candidate (the polyline highlighted in yellow)

Maps are produced respectively for the original male/female population counts and associated CVs, and the aggregated male/female population counts and associated

CVs (Figure 36 and Figure 37). Aggregation seeds in the original data are highlighted on the maps (Figure 36A, Figure 36C, Figure 37A and Figure 37C), and the newly generated units are highlighted on the maps of aggregated data (Figure 36B, Figure 36D, Figure 37B and Figure 37D). In total, the number of areal units was reduced from 94 to 82 with 12 enumeration units being merged to the 12 aggregation seeds. Every seed were merged with one surrounding unit, and such aggregations were sufficient to keep al CV values below the 9.5 threshold. In order to observe the changes of estimates and CVs across different maps, I use the same classification for the corresponding maps. The maps of estimates were made using Jenks natural breaks method based on the aggregated data. Since the range of the aggregated estimate is larger than the original estimate, the classes made by the aggregated estimate can include all values of the original estimate. Similarly, the maps of CV values were made using equal interval method based on the original data.

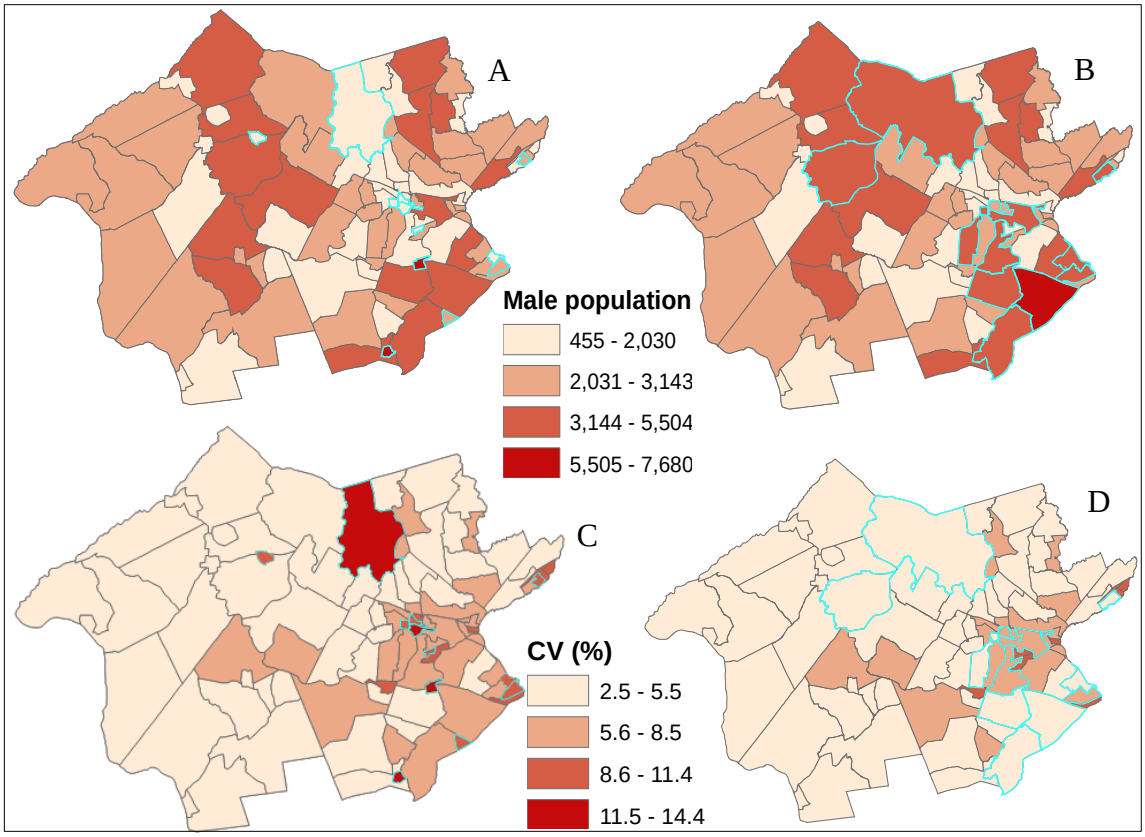

Figure 36 Maps of estimates and CV of the male population counts before and after aggregation (A: male population counts before aggregation with seeds highlighted in cyan; B: male population counts after aggregation with new areal units highlighted in cyan; C: CVs of male population counts before aggregation with seeds highlighted in cyan; D: CVs of male population counts after aggregation with new areal units highlighted in cyan)

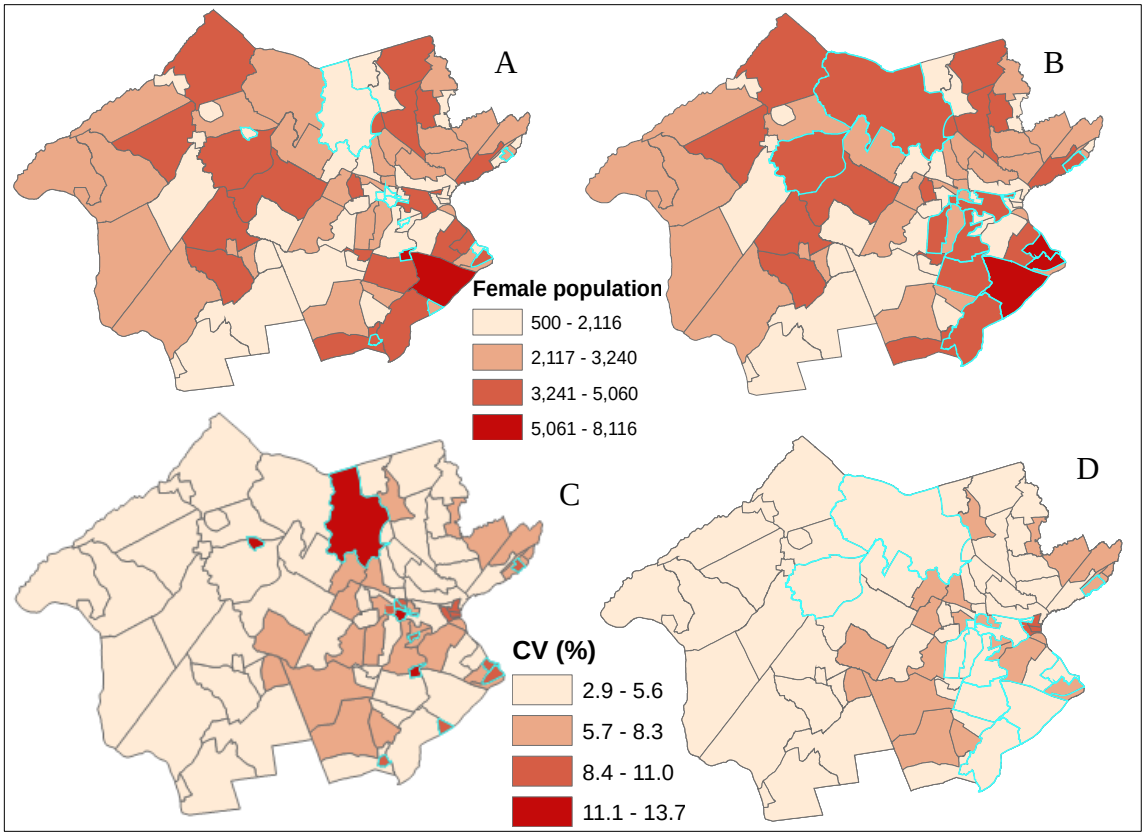

Figure 37 Maps of estimates and CV of the female population counts before and after aggregation (A: female population counts before aggregation with seeds highlighted in cyan; B: female population counts after aggregation with new areal units highlighted in cyan; C: CVs of female population counts before aggregation with seeds highlighted in cyan; D: CVs of female population counts after aggregation with new areal units highlighted in cyan)

Based on those maps, we found that the CVs of the aggregated observations were reduced from their original ones significantly. Several original units with CVs assigned to the highest or the second highest classes on the east side of the study area fell into the lower classes after aggregation (Figure 36C, Figure 36D, Figure 37C and Figure 37D). As shown in Table 7, while the maximum CVs of the original male and female population counts were 14.4 and 13.7 percent respectively, the corresponding new CV values are 9.6 and 8.6 percent. The means decreased slightly, but the SDs decreased

significantly since the large CV values (the ones larger than 9.5 percent) were removed from the datasets. As expected, the minimum CV values do not change.

Regarding the aggregated estimates, the new population counts were sums of the counts of all observations being merged together. The minimum, maximum, mean and SD of the data with new estimates become larger, but the overall error levels have diminished (Table 7). We understand that the changes in the summary statistics are specific to the count variables we have chosen in this demonstration. When other types of variables, such as ratio variables, certain summary statistics like the maximum may not increase after aggregation. The spatial pattern changed to some extent, especially in the local area involving aggregation (Figure 36A, Figure 36B, Figure 37A and Figure 37B). Some high-low clusters changed to high-high clusters, and some low-low clusters change to high-low clusters. This is an unavoidable problem (i.e. MAUP) caused by aggregation. However, the pattern in areas not involved in the aggregation did not change too much. Also, besides some new areal units were assigned into high-value classes, the global pattern of the aggregated data presented by the current classification (Figure 36B and Figure 37B) is very similar to the global pattern of the original data (Figure 36A and Figure 37A). The constraint that aggregation is only performed on enumeration units with large error does help to reserve the spatial variation in the original data for most units.

Table 7 Selected statistics for the estimates and errors of the male and female population before and after aggregation

	Statistics	*Original*		*Aggregated*	
		male	female	male	female
Estimate	Min	188	226	455	500
	Max	5,220	5,863	7,680	8,116
	Mean	2,348	2,436	2,691	2,792
	SD	905	1,008	1,126	1,234
CV (%)	Max	14.4	13.7	9.6	8.6
	Min	2.5	2.9	2.5	2.9
	Mean	6.2	5.9	5.4	5.2
	SD	2.4	2.3	1.7	1.5
	Units	94		82	

While the aggregated data have more reliable estimates, we may still ask: how does the improvement on estimate reliability contribute to the result derived from mapping or data analysis? In order to the answer this question, we select one variable, the male population counts, to make two maps simply based on the Jenks natural breaks method using the original and aggregated data (Figure 38). The separability measure was used to evaluate the reliability of the map classifications (Table 8). Note that using map classification to demonstrate the usability of aggregated data does not mean that the purpose of aggregation is to make reliable classification. Instead, we just take mapping as an example to demonstrate the influence of aggregation results on data analysis in GIS since mapping is actually an analysis process (Sun et al. 2014).

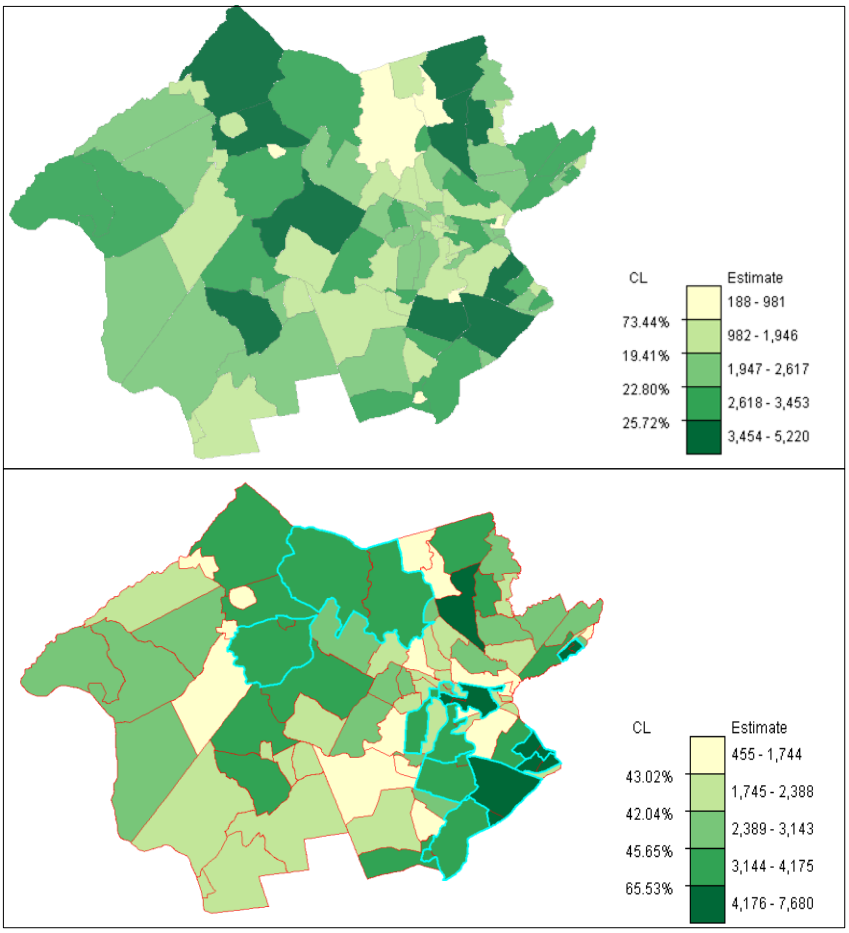

Figure 38 Maps using the original and aggregated estimates with separability level associated with each class break (Upper: the original male population count estimates; Lower: the aggregated male population estimates)

As shown in Table 8, in the map of the original estimates, except the high separability level for the break between the first and second classes, the separability levels of the classification are around 20 percent. All these levels were improved dramatically to at least 42 percent in the classification for the aggregated estimates. The high separability level approaches to 66 percent. The average separability was improved from 35 percent to 49. Therefore, we are more confident with the classes created based

on the aggregated data, and the spatial pattern presented through the differences among classes is more reliable (Figure 38 lower).

Table 8 Separability levels of class breaks in the map of aggregated estimate

Classes	*Original*	*Aggregated*
1-2	73	43
2-3	19	42
3-4	23	46
4-5	26	66
Mean	35	49

In summary, we designed a series of demonstrations to show the effectiveness of the proposed mapping and aggregation methods. Two population datasets, the ACS and national mortality data, are selected for the demonstrations. The classification method that maximizes class separability helps to produce separable classes very effectively and performs much better than traditional classification methods. However, the method may create highly unbalanced classifications which may not be too useful to detect spatial patterns. On the other hand, the heuristic mapping method, which takes into account multiple criteria in the mapping process, is more flexible in considering multiple criteria, including separability to create likely more useful maps. The aggregation procedure was demonstrated to be able to produce more reliable estimates. The spatial pattern inherent in the original data was changed, but the process preserves the local variations to the largest extent.

CHAPTER SIX SUMMARY AND DISCCUSION

6.1 Summary

In this dissertation, we first proposed a class-separability metric and related heuristic mapping methods based on the separability metric. These methods were developed with the intent to augment our capabilities to map spatial data incorporating data quality information. Different from the traditional classification methods which divide observations into different groups based only on the distribution of estimates, the proposed methods consider both estimate values and the error of estimates in the classification process.

One of the methods is to determine map class breaks that seek to maximize the confidence that values in adjacent classes are statistically different. When more classes are formed, values between classes become less different. Therefore, cartographers or map users have to determine the trade-off between separability levels and number of classes. However, since the above classification method incorporates only two criteria in determining class breaks (the separability between classes and the number of classes), excluding the distribution of estimates across classes, the usefulness of resultant maps in revealing spatial patterns may be limited. Therefore, to make the maps more effective to show spatial patterns, other criteria, such as the homogeneity within classes (e.g. Smith 1986), spatial compactness of units assigned to the same class (e.g. Cromley 1996), and underlying spatial structure (autocorrelation) in the data (e.g. Murray and Shyy 2000)

may need to be considered. Accordingly, another advanced heuristic classification method was proposed to take into account multiple criteria in determining class breaks. The roles of and the need for different criteria may be case-dependent, and the objectives of the map should dictate which criteria should be included. Depending upon the objective, it may also be desirable to balance the numbers of observations among classes. The two proposed mapping methods are both data driven. If estimates have relatively large errors, then the separability levels will generally be low, and the probabilities of putting estimates in classes that they should belong to, regardless of which classification method is adopted, will be relatively low.

If maps are made using conventional methods, such as Jenks natural breaks, the class-separability metric can be used to evaluate the extent that values between classes are statistically different and the reliability of the spatial pattern reflected by the differences between classes. Then a map legend, in which a confidence level of separability is associated with each class break, was designed to indicate to what extent the two adjacent classes are statistically separable.

Despite the methods proposed above, maps and analysis results cannot be more reliable than what provided by the quality of input data. We, therefore, took a slightly different approach to reduce error from data directly. Spatial aggregation is one method which data users can use to reduce error in data by creating new areal units with more reliable estimates, but at the expenses of changing the spatial configuration of the original data with larger areal units and introducing bias to the new estimates. To minimize the possible changes in the spatial configuration, our proposed procedure will aggregate only

areal units (seeds) with error levels so large that they cannot be acceptable by data users. Data users decide the threshold of acceptable error size based on the requirement of the particular applications. Our procedure helps find out the areal units that potentially can be merged with aggregation seeds to derive new estimates with acceptable error levels. Human intelligence is required in determining the best choices of areal unit(s) from the candidate pool created by our procedure. Similar to the heuristic mapping process, multiple criteria such as spatial compactness and spatial autocorrelation, etc., can be included in the aggregation process to guide data users on selecting the "best" aggregation configuration to produce more usable data. Since the areal units to be merged are decided by data users, the aggregation configuration may not be optimal (according to certain criteria), but the aggregated data will be, at least, appropriate or usable for the particular purpose.

In addition, as multiple criteria are involved in determining the map classification and aggregation scheme, a rather flexible framework, including effective graphics and computer-user interaction, is needed to evaluate all reasonable options available to data users and the trade-offs between different criteria. Therefore, we developed a suite of interactive visual tools to facilitate the proposed heuristic mapping methods and aggregation procedure. To some extent, the prototype toolset can also be customized to serve the other data analysis involving multi-dimensional information.

Two popular data sources, the ACS data and the national mortality data, which capture the characteristics of socioeconomic and public health aspects of our society, are used to demonstrate the effectiveness of the proposed methods and associated tools. In

real world practices, some more data sets are more applicable. Theoretically, most data collected by sampling surveys, including estimates and errors measure by SE, MOE, and CV, etc., may utilize our proposed methods to incorporate attribute accuracy into the mapping process.

The methods introduced in this dissertation were developed using statistical concepts. Therefore, users should have statistical knowledge to understand and use the proposed methods. Users should at the minimum understand the concepts related to standard error and the difference of means test. Training will be needed for the general public to acquire related knowledge before they use the proposed methods to handle the attribute accuracy information in spatial data.

6.2 Future Work

Several issues need to be addressed in the future. The first issue is related to the fundamental of the class separability metric. To determine the separability of classes (i.e. CL of difference between classes), we actually test if estimates between two classes are significantly different. Every pair of estimates is compared using t-test respectively and the separability is the minimum CL of difference between a pair of estimate. However, comparing multiple pairs of estimates from two groups may lead to a multiple comparison problem in statistics: the CL of difference derived from comparing the two groups as a whole directly is not equal to the one derived from comparing every pair of observations in the two groups (Hochberg and Tamhane 1987). Type I error, i.e. incorrectly reject the null hypothesis that two classes are significantly different while the classes are actually not, is more likely to occur than one considers the class a whole. In

other words, the significance of difference test result will be more confident if considering every class as a whole in the comparison. Therefore, the current class separability measure should be modified by utilizing multiple comparison methods, such as the least significant difference method (Williams and Abdi 2010), rather than comparing each pair of estimates in two classes.

In addition, when performing t-test, the assumption is that the sample means (i.e. estimates) are normally distributed. While the estimates of most variables are likely normally distributed as long as the sample size is "large", this may not be true for some variables like income. For example, the distribution of income is more closely to log normal than normal (Mitzenmacher 2004). The mismatch between the t-test assumption and estimate distribution may be addressed by transforming data into normal distribution, or conducting the non-parametric test (e.g. the Mann-Wilcoxon test).

Second, the heuristic mapping method that considers multiple criteria is based on a particular standard classification method, and the designed star plot can only display the potential classification schemes made by this method. The mapping method can be improved by more exhaustive "search". The star plot may incorporate potential schemes made by different classification methods to enable cross-method comparison. Furthermore, map makers are required to select a classification method from potential schemes with different number of classes. To some extent, the heuristic process is to incorporate multiple criteria to guide the selection of the "best" number of classes that can create the scheme with acceptable separability among classes and moderate values on the other criteria. The class breaks are still determined by standard classification methods.

To overcome the limitation, we may need to design a more flexible classification procedure to search for the breaks by evaluating the trade-off relationship among different criteria. For large dataset, such search process will lead to intensive computation, and therefore, high performance computing infrastructure maybe needed to ensure the efficiency of execution.

Third, the process involving human intelligence to determine the best candidate to be merged with the aggregation seed does not guarantee that the aggregation scheme is optimal. Searching candidate and merging the best candidate with the aggregation seed were executed sequentially starting from the seed with the largest CV. Candidates merged with seeds earlier in the process may be better for seeds later in the process to achieve closer to a global optimal situation. Therefore, optimization algorithms should be integrated with the current human-centric method to increase the likelihood of achieving the optimal combination of candidates and seeds. To approach this, we may include an extract step to check if swapping candidates to different seeds will improve the "optimality" after manually selecting the "best" options for all seeds. In addition, the methods of estimating the new estimates and errors after aggregation have be developed for the other types of statistics like proportions, means and median besides counts.

In addition, tests are needed to evaluate the difficulties in comprehending the proposed methods and the usability of the supporting tools. Tests on different target user groups, such as scientists, cartographers, economists and census officials, are desirable. They may be required to perform different related tasks, such as to produce map or perform aggregation with the tools, and then provide feedback on their experience (e.g.

how easy or difficult to perform these tasks). Metrics also need to be defined to evaluate the reliability of maps or data produced by users using these tools.

REFERENCES

Anselin, L., E. Griffiths and G. Tita. 2008. Crime Mapping and Hot Spot Analysis. In: R. Wortley and L. Mazerolle (eds.), *Environmental Criminology and Crime Analysis*: 97-116.

Anselin, L. 1994. Exploratory Spatial Data Analysis and Geographic Information Systems. In M. Painho (eds.), *New Tools for Spatial Analysis, Eurostat*, Luxembourg: 45-54.

Armhein, C. 1995. Searching for the elusive aggregation effect: Evidence from statistical simulations. *Environment & Planning A*, 27(1): 105.

Banda, J. 2003. Current Status of Social Statistics: An overview of Issues and Concerns. *United Nations Expert Group Meeting in collaboration with the Siena Group on Social Statistics*, New York, 6-9 May 2003.

Beard, M.K. 1991. Position Statement on the visualization of data quality Coleaders. In M.K. Beard and Buttenfield B. (eds.), *NCGIA Specialist Meeting of 7 - Visualisation of the Quality of Spatial Data*, June 8-12 Castine, ME, USA.

Beard, M.K., B. Buttenfield and S. Clapham. 1991. *NCGIA NCGIA Research Initiative 7 Visualisation of the Quality of Spatial Data*. Scientific report for the specialist meeting, 8-12 June 1991, Castine, Maine.

Bell, W. and D. Wilcox. 1991. The effect of sampling error on the time series behavior of consumption data. *Bureau of the Census Statistical research Division Report Series*, CENSUS/SRD/RR-91/02.

Brewer, C. A. and Suchan, T. A. 2001. Mapping Census 2000: The Geography of U.S. Diversity, *Census Special Report, Series CENSR/01-1*, US Government Printing Ofice, Washington DC.

Brewer, C.A. and L. Pickle. 2002. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4): 662-81.

Briant, A., P.P. Combes and M. Lafourcade. 2008. Dots to Boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations?. *CEPR Discussion Paper Series.*

Burrough, P.A. and R.A. McDonnell. 1998. *Principles of geographical information systems for land resources assessment*, Oxford University Press, USA.

Burt, J.E. and G.M. Barber. 1995. *Elementary Statistics for Geographers: Second Edition*, The Guilford Press, NYC, USA.

Buttenfield, B.P. 1991. Visualizing Cartographic Metadata. *Proceedings, NCGIA Specialist Meeting on Visualizing the Quality of Spatial Data*, June 1991, Castine, Maine: 19-28.

Carrasco, P.C. 2009. Nugget effect, artificial or natural? *Paper in Fourth World Conference on Sampling & Blending*, the Southern African Institute of Mining and Metallurgy, 2009.

Carr, D. B. and L. W. Pickle. 2010. *Visualizing Data Patterns with Micromaps*, Taylor & Francis Group (CRC Press), Boca Raton FL.

Carr, D.B., D. White and A.M. MacEachren. 2005. Conditioned choropleth maps and hypothesis generation. *Annals of the Association of American Geographers,* 95: 32-53.

Chambers. J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey. 1983. *Graphical methods for data analysis*, New York : Champman & Hall.

Clementini, E, P. Di Felice and P. van Oosterom. 1993. A small set of formal topological relationship suitable for end-user interaction. in D. Abel and B.C. Ooi (eds.), *Third International Symposium on Large Spatial Databases. SSD '93*. *Lecture Notes in Computer Science 692*, pp 277-295, Springer-Verlag, New York, NY.

Cleveland, W. S. 1994. *The elements of graphing data*. Monterey, CA: Wadsworth advanced books and software.

Cockings, S. and D. Martin. 2005. Zone design for environment and health studies using pre-aggregated data. *Social Science & Medicine*, 60: 2729-2742.

Cressie, N. 1992. Statistics for spatial data. *Terra Nova*, 4(5): pp 613-617.

Cromley, R. G. 1996. A comparison of optimal classification strategies for choroplethic displays of spatially aggregated data. *International Journal of Geographic Information Science* 10 (4): 405 - 24.

Datta, D., J. Malczewski and J.R. Figueira. 2012. Spatial aggregation and compactness of census areas with a multiobjective genetic algorithm: a case study in Canada. *Environment and Planning B: Planning and Design*, 39(2) 376 – 392.

Devillers, R., Y. Bédard and R. Jeansoulin. 2005. Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use within GIS. *Photogrammetric Engineering and Remote Sensing* 71: 205–15.

de Vise, D. 2010. Rich mix of academic backgrounds emerges in D.C. area census data. *Washington post news*, October 15, 2010.

Dobson, M.W. 1973. Choropleth maps without class intervals?: A comment. *Geographical Analysis*, 5(3):358-360.

Edwards, L.D. and E.S. Nelson. 2001. Visualizing Data Certainty: A Case Study Using Graduated Circle Maps. *Cartography perspective* 38: 19-36.

Ehlschlaeger, C.R., A.M. Shortridge and M.F. Goodchild. 1997. Visualizing spatial data uncertainty using animation. *Computers and Geosciences*, 23(4): 387–395.

Evans, B. J. 1997a. Dynamic display of spatial data-reliability: does it benefit the map user?. *Computers and Geosciences*, 24(1): 1–14.

Evans, I. S. 1977b. The selection of class intervals. *Transactions of the Institute of British Geographers*, New Series 2(1): 98-124.

FGDC. 1998. *Content Standard for Digital Geospatial Metadata (version 2.0)*. Available from: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf (Accessed by Aug 2, 2012).

Folch, D. and S. E. Spielman. 2014. Identifying regions based on flexible user-defined constraints. *International Journal of Geographical Information Science*, 28(1): 164-184.

Friedrich, G.W. 2000. Sampling theory methods of inquiry syllabus: 514. Available from: http://comminfo.rutgers.edu/~gusf/sampling.html (Access by Aug 2, 2012).

Frolov, Y. 1975. Measuring of shape of geographical phenomena: a history of the issue. *Soviet Geography: Review and Translation*, 16: 676—687.

Gahegan, M., M. Takatsuka, M. Wheeler and F. Hardisty. 2000. GeoVISTA Studio: a geocomputational workbench. *Proceedings of the Fifth International Conference on GeoComputation*, Chatham, London, August 23–25, 2000.

Gahegan, M., M. Takatsuka, M. Wheeler, F. Hardisty. 2002. Introducing GeoVISTA Studio: An integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems*, 26(4): 267–292.

Gan, E. and W. Shi. 2002. Error Metadata Management System. In Shi, W., P. F. Fisher, and M. F. Goodchild (Eds.), *Spatial Data Quality*, London: Taylor and Francis: 251–66.

Ghouse, Z.M., and M. Duckham. 2009. Integrated Storage and Querying of Spatially Varying Data Quality Information in a Relational Spatial Database. *Transactions in GIS*, 13(1): 25–42.

Glover, F. 1977. Heuristics for integer programming using surrogate constraints, decision. *Science*, 8: 156-166.

Glover, F. 1986. Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research* 13(5): 533-549.

Goodchild, M.F. 1995. Preface. In W. Shi, P. F. Fisher, and M.F. Goodchild (eds). *Spatial Data Quality*, CRC Press, FL, USA.

Goodchild, M.F. 2003. Models for uncertainty in area-class maps. In W. Shi, M. F. Goodchild, and P. F. Fisher (eds). *Proceedings of the Second International Symposium on Spatial Data Quality*. Hong Kong: Hong Kong Polytechnic University, pp. 1–9.

Griffith, D. 1991. Data quality and visualization: a position paper, in M. Beard, B. Buttenfield and S. Clapham (eds.), *NCGIA Research Initiative 7: Visualization of Spatial Data Quality, Techinical Paper No. 91-26*, NCGIA, C77-C85.

Guo, J.Y., G. Trinidad and N. Smith. 2001. MOZART: a Multi-objective Zoning and AggRegation Tool. *Paper presented at the TRB 80th Annual Meeting, Transportation Research Board*, Washington, DC.

Haining, R., S. Wise and J. Ma. 1998. Exploratory spatial data analysis in a geographic information system environment. *The Statistician*, 47 (Part 3): 457 -469.

Hengl, T., D. J. J. Walvoort, A. Brown, and D. G. Rossiter. 2004. A double continuous approach to visualization and analysis of categorical maps. *International Journal of Geographical Information Science* 18 (2): 183-202.

Hess, S.W., J.B. Weaver, J.N. Siegfeldt, J.N. Whelan and P.A. Zhitlau. 1965. Nonpartisan political districting by computer. *Operations Research*, 13: 998-1006.

Heuvelink, G.B.M. and P.A. Burrough. 1989. Propagation of Errors in Spatial Modeling with GIS. *International Journal of Geographical Information Science*, 3(4): 303-322.

Heuvelink, G.B.M. and P.A. Burrough. 2002. Developments in statistical approaches to spatial uncertainty and its propagation. *International Journal of Geographical Information Science*, 16: 111–3.

Hochberg, Yosef, and Ajit C. Tamhane. 1987. *Multiple comparison procedures*. John Wiley & Sons, Inc.

Horn, M.E.T. 1995. Solution techniques for large regional partitioning problems. *Geographical Analysis*, 27: 230-248.

Hope, S. and G.J. Hunter. 2007. Testing the effects of thematic uncertainty on spatial decision-making. *Cartography and Geographic Information Science*, 34(3):199-214.

Howard, D. and A.M. MacEachren. 1996. Interface design for geographic visualization: tools for representing reliability. *Cartography and Geographic Information Systems*, 23:59-77.

Hunter, G.J. and M.F. Goodchild. 1996. A new model for handling vector data uncertainty in geographic information systems. *Journal of the Urban and Regional Information Systems Association*, 8(1): 51–57.

Jakobsson, A. 2002. Data quality and quality management- examples of quality evaluation procedures and quality management in European National Mapping Agencies. In W. Shi, P. F. Fisher, and M.F. Goodchild (eds). *Spatial Data Quality*, CRC Press, FL, USA: 216-229.

Jenks, G. and M. Coulson. 1963. Class intervals for statistical maps. *International Yearbook of Cartography*, 3:119-133.

Jenks, G.F. 1977. Optimal data classification for choropleth maps. University of Kansas, Department of Geography Occasional Paper No. 2, Lawrence, Kansas.

King, K. 2011. Aggregating Estimates and Calculating Margins of Errors. *Presentation at the Joint FSCPE/CIC/SDC Meeting*, February 16, 2011.

Kobus, D., S. Proctor and S. Holste. 2001. Effects of expemaking. *International Journal of Industrial Ergonomics*, 28: 275-90.

Koua, E.L. and M. Kraak. 2008. An integrated exploratory geovisualization environment based on self-organizing map. In Agarwal P., Skupin A.,(Eds.), *Self-Organising Maps: Applications in Geographic Information Science*, Wiley: 45-66.

Kohonen, T., M. R. Schroeder and T. S. Huang. 2001. *Self-Organizing Maps 3rd*, Springer-Verlag New York, Inc. Secaucus, NJ, USA.

Leitner, M. and B.P. Buttenfield. 2000. Guidelines for the Display of Attribute Certainty. *Cartography and GIS*. 27(1): 3-14.

Williams, L. J. and H. Abdi. 2010. Fisher's least significance difference (LSD) test. *Encyclopedia of Research Design*. Thousand Oaks: 491-494.

Li, W., R. Church and M. Goodchild. 2014. An extendable heuristic framework to solve the p-compact-regions problem for urban economic modeling. *Computers, Environment and Urban Systems* 43 (2014) 1–13.

Li, W., M.F. Goodchild and R. Church. 2013. An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *International Journal of Geographical Information Science*, 27 (6): 1227–1250.

Longley, P.A., M.F. Goodchild, D.J. Maguire and D.W. Rhind. 2001. *Geographic Information Systems and Science*, Wiley, NYC.

MacDonald, H. 2006. The American community survey: Warmer (more current), but fuzzier (less precise) than the decennial census. *Journal of the American Planning Association* 72(4): 491-503.

MacEachren, A.M. 1985. Compactness of geographic shape: comparison and evaluation of measures. *Geografiska Annaler, Series B*, 67: 53 - 67.

MacEachren, A.M. 1992. Visualizing uncertain information. Cartographic Perspectives *Cartographic perspectives*, 13 Fall: 10-19.

MacEachren, A.M., D. Howard, M. von Wyss, D. Askov and T. Taormino. 1993. Visualising the health of Chesapeake Bay: an uncertain endeavor. In *Proceedings GIS/LIS 93 Minneapolis*: 449–58.

McGranaghan, M. 1993. A cartographic view of spatial data quality. *Cartographica*, 30(2): 8-19.

MacEachren, A.M., Thacher, J., & Reeves, C. 1994. Some Truth with Maps: A Primer on Symbolization and Design. Association of American Geographers, Washington, D.C.

MacEachren, A.M., C.A. Brewer, and L.W. Pickle. 1998. Visualizing georeferenced data: Representing reliability of health statistics. *Environment & Planning A*, 30: 1547-61.

MacEachren, A.M., M. Wachowicz, R. Edsall, D. Haug and R. Masters. 1999. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13: 311-334.

MacEachren, A.M., A. Robinson, S. Hopper, S. Gardner, R. Murray and M. Gahegan. 2005. Visualizing Geospatial Information Uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32:139-160.

Macmillan, W., T. Pierce. 1994. Optimization modeling in a GIS framework: the problem of political redistricting, In S. Fotheringham and P. Rogerson (eds.), *Spatial Analysis and GIS*, Taylor and Francis, London: 221 - 246.

Makuc D, Haglund B, Ingram D, Kleinman J, Feldman J. Health Services Areas for the United States. 1991. *National Center for Health Statistics, Series 2, No. 112.*

Martin, D. 1998. Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Science*, 12(7), 673–685.

Martin, D., A. Nolan and M. Tranmer. 2001. The application of zone design methodology to the 2001 UK Census. *Environment and Planning A*, 33(11): 1949–1962.

Martin, D. 2003. Developing the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 17(2), 181–196.

Miller, H. J. and J. Han. 2001. *Geographic Data Mining and Knowledge Discovery*, London and New York, Taylor & Francis: 3-32.

Mitzenmacher, M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet mathematics* 1, no. 2: 226-251.

Ghouse, Z. M. 2008. Modeling spatial variation of data quality in databases. *PhD thesis, Faculty of Engineering, Geomatics,* The University of Melbourne.

Murray, A.T. and T. Shyy. 2000. Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science* 14: 649-667.

Esri. 2004. Louisville water users GIS data ReViewer to check data quality. http://www.esri.com/software/arcgis/extensions/arcgis-data-reviewer/reviewer_louisville.pdf (last accessed by May 31, 2014).

Monmonier, M. 1989. Geographic brushing: Enhancing exploratory analysis of the scatter plot matrix. *Geographical Analysis*, 21(1): 81-84.

Monmonier, M. S. 1972. Contiguity-based class-interval selection: a method for simplifying patterns on statistical maps. *The Geographical Review,* 62 (2): 203-228.

NIST (National Institute of Standards and Technology). 1992. *Federal Information Processing Standard, Publication 173* (Spatial Data Transfer Standard): U.S. Department of Commerce, Washington, D.C., variously paged.

Olson, J.M. 1981. Spectrally encoded two-variable maps. *Annals of the Association of American Geographers*, 71(2): 259-76.

Openshaw, S. and L. Rao. 1995. Algorithms for reengineering 1991 census geography. *Environment and Planning A*, 27: 425-446.

Openshaw, S. 1977. A geographical solution to scale and aggregation problems in regionbuilding, partitioning, and spatial modeling. *Transactions of the Institute of British Geographers*, New Series: 2459-472.

Openshaw, S. 1978. An optimal zoning approach to the study of spatially aggregated data. In I. Masser and P.J.B. Brown (eds.), *Spatial Representation and Spatial Interaction*, Martinus Nijhoff, Leiden: 93-113.

Openshaw, S. and P. Taylor. 1979. A million or so correlation coefficients: three experiments on the modifiable area unit problem. In N. Wrigley (eds), *Statistical Applications in the Spatial Sciences*, London. Pion:127-144 .

Openshaw, S. 1983 The modifiable areal unit problem. *Concepts and Techniques in Modern Geography No. 38*. Geobooks, Norwich, England.

Openshaw, S. 1988. Building an automated modeling system to explore a universe of spatial interaction models. *Geographical Analysis*, 20: 31-46.

Openshaw, S., S. Alvanides. 1999. Applying geocomputation to the analysis of spatial distributions P. Longley, M. Goodchild, D. Maguire, D. Rhind (Eds.), *Geographical Information Systems: Principles, Techniques, Applications and Management*, Wiley, Chichester: 267–282.

Ostensen, O. and P.C. Smits. 2002. *ISO/TC211: Standardisation of Geographic Information and Geo-Informatics*. Available from: http://www.erc.msstate.edu/committees/GRSS-DAD/igarss0206.pdf (Accessed by Aug 2, 2012).

Pickle, L.W., M. Mungiole, G.K. Jones and A.A. White. 1996. *Atlas of United States Mortality*. Hyattsville, MD:National Center for Health Statistics.

Qiu, J. and G. J. Hunter, 2002, A GIS with the Capacity for Managing Data Quality Information. In Shi, W., P. F. Fisher, and M. F. Goodchild (Eds.), *Spatial Data Quality*, London: Taylor and Francis: 230–250.

Ralphs, M. and L. Ang. 2009. Optimised geographies for data reporting: zone design tools for census output geographies. WP 09-0, Statistics New Zealand, Wellington.

Ripley, B.D. 1981. *Spatial Statistics*, New York: John Wiley Sons.

Sammons, R. 1977. A simplistic approach to the redistricting problem. In I. Masser and P. Brown (eds.), *Spatial Representation and Spatial Interaction*, Leiden: Martinus Nijhoff.

Schultz, G. M. 1961. An Experiment in Selecting Value Scales for Statistical Distribution Maps, *Surveying and Mapping*, 21: 224-230.

Shirabe, T. 2009. Districting modeling with exact contiguity constraints. *Environment and Planning B: Planning and Design*, 36: 1053-1066.

Siska, P.P., and I. Hung. 2001a. Propagation of errors in spatial analysis. *Papers and Proceedings of the Applied Geography Conferences*, University of North Texas 24: 284-290.

Siska, P.P.,and I. Hung. 2001b. Assessment of Kriging Accuracy in the GIS Environment, Available from: http://proceedings.esri.com/library/userconf/proc01/professional/papers/pap280/p280.htm (Accessed by Aug 2, 2012).

Slocum, T.A., R.B McMaster, F.C. Kessler and H.H. Howard. 2003. *Thematic Cartography and Geographic Visualization*. New York: Prentice Hall.

Smith, R. M. 1986. Comparing traditional methods for selecting class intervals on choropleth maps. *The Professional Georgrapher* 38 (1): 62-67.

Sombroek, W. and A. Carvalho. 2000. Macro- and Micro Ecological-economic Zoning in the Amazon Region History, First Results, Lessons Learnt and Research Needs. In *German-Brazilian Workshop on Neotropical Ecosystems – Achievements and Prospects of Cooperative Research*. Hamburg, September 3-8, 2000.

Spielman, S.E., D. Folch and N. Nagle. 2014. Patterns and causes of uncertainty in the American Community Survey. *Applied Geography* 46: 147-157. doi: 10.1016/j.apgeog.2013.11.002

Stegna, L., and F. Csillag. 1987. Statistical determination of class intervals of maps. *The Cartographic Journal* 24 (2): 142-46.

Su Y., L. Yang and Z. Jin. 2007. Evaluating Spatial Data Quality in GIS Database. WiCom 2007. *International Conference on Wireless Communications, Networking and Mobile Computing*, Shanghai, Sep 2007: 5967-5970.

Sun, M., and D. W. S. Wong. 2010. Incorporating Data Quality Information in Mapping the American Community Survey (ACS) Data. *Cartography and Geographic Information Science*, 37(4): 285-300.

Sun, M., Wong, D. W. and Kronenfeld, B. J. 2014. A Classification Method for Choropleth Maps Incorporating Data Reliability Information. *The Professional Geographer*, doi:10.1080/00330124.2014.888627

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* 2013 1(1): 3-16.

Thomson, J., B. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel. 2005. Typology for visualizing uncertainty. Conference on Visualization and Data Analysis, *Part of the IS&T/SPIE Symposium on Electronic Imaging* 2005.

Tobler, W.R. 1973. Choropleth maps without class intervals?. *Geographical Analysis*, 5(3): 262-265.

Town, R. J., D. Wholey, R. Feldman and L. R. Burns. 2007. Revisiting the Relationship between Managed Care and Hospital Consolidation. *Health Serv Res*. Feb 2007; 42(1 Pt 1): 219–238, doi: 10.1111/j.1475-6773.2006.00601.x.

Tukey, J. 1977. *Exploratory Data Analysis*, 1st edition. Pearson.

.
The United Nations, Department of Economic and Social Affairs Statistics Division, Demographic and Social Statistics Branch. 2004. *United Nations Demographic Yearbook review*. http://unstats.un.org/unsd/demographic/products/dyb/techreport/mortality.pdf (last accessed by May 14, 2014).

Unwin, D. 1995. Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography*, 19: 549-558.

U.S. Census Bureau. 2008. A compass for understanding and using American Community Survey data – What general data users need to know. Available from: http://www.census.gov/acs/www/Downloads/handbooks/ACSGeneralHandbook.pdf (Accessed by Aug 2, 2012).

U.S. Census Bureau. 2012a. Surveys, Available from: http://www.census.gov/aboutus/surveys.html (Accessed by Aug 2, 2012).

U.S. Census Bureau. 2012b. Available from: www.census.gov/indicator/qss/qsstechdoc.pdf (Accessed by Aug 2, 2012).

Vanegas, P., D. Cattrysse and J. van Orshoven. 2010. Compactness in spatial decision support: a literature review. *Lecture Notes in Computer Science*, 6016: 414-429.

Veregin, H. 1999. Data quality parameters. In. P.A. Longley, M.F. Goodchild, D.J. Maguire and D. W. Rhind (eds), *Geographical Information Systems: Principles, Techniques, Management and Applications*, John Wiley and Sons, USA.

Xia, H. and Carlin, B. P. 1998. Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine*, 17:2025–2043.

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): 121-44.

Xiao, N. and M.P. Armstrong. 2006. ChoroWare: a software toolkit for choropleth map classification. *Geographical Analysis*, 38(1), 102–121.

Wong, D.W.S. and C.V. Wu. 1996 Spatial metadata and GIS for decision support. In *Proceedings of the Twenty-ninth Annual Hawaii International Conference on System Sciences*, Honolulu, Hawaii.

Wong, D. W. S. 1996. Aggregation Effects in Geo-Referenced Data. In D. Griffths (eds.), *Advanced Spatial Statistics*, Boca Raton, FL: CRC Press: 83–106.

Wong, D. W. S. 2003. Implementing spatial segregation measures in GIS. *Computers, Environment and Urban Systems*, 27, 53-70.

Wong, D and J. Lee. 2005. *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. John Wiley and Sons, Inc., USA.

Wong, P.C. and J. Thomas. 2004. Visual Analytics. In *IEEE Computer Graphics and Applications*, 24(5) Sept.-Oct.2004: 20–21.

Wright, John Kirtland. 1938. *Notes on Statistical Mapping: With Special Reference to the Mapping of Population Phenomena*. American Geographical Society.

BIOGRAPHY

Min Sun received her Bachelor of Science in Geographic Information Science from Nanjing Normal University, China in 2008. She began her study in the Ph.D. program at George Mason University from 2008, Fairfax, VA 22031. E-mail: msun@gmu.edu. Her research interests include measuring attribute uncertainty in spatial data, developing visual analytics to support spatial data exploration and WebGIS.