Visual Analytics for

Spatiotemporal Cluster Analysis

by

Yifan Zhang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2016 by the
Graduate Supervisory Committee:

Ross Maciejewski, Chair
Elizabeth Mack
Huan Liu
Hasan Davulcu

ARIZONA STATE UNIVERSITY

May 2016

ABSTRACT

Traditionally, visualization is one of the most important and commonly used methods of generating insight into large scale data. Particularly for spatiotemporal data, the translation of such data into a visual form allows users to quickly see patterns, explore summaries and relate domain knowledge about underlying geographical phenomena that would not be apparent in tabular form. However, several critical challenges arise when visualizing and exploring these large spatiotemporal datasets. While, the underlying geographical component of the data lends itself well to univariate visualization in the form of traditional cartographic representations (e.g., choropleth, isopleth, dasymetric maps), as the data becomes multivariate, cartographic representations become more complex. To simplify the visual representations, analytical methods such as clustering and feature extraction are often applied as part of the classification phase. The automatic classification can then be rendered onto a map; however, one common issue in data classification is that items near a classification boundary are often mislabeled.

This thesis explores methods to augment the automated spatial classification by utilizing interactive machine learning as part of the cluster creation step. First, this thesis explores the design space for spatiotemporal analysis through the development of a comprehensive data wrangling and exploratory data analysis platform. Second, this system is augmented with a novel method for evaluating the visual impact of edge cases for multivariate geographic projections. Finally, system features and functionality are demonstrated through a series of case studies, with key features including similarity analysis, multivariate clustering, and novel visual support for cluster comparison.

To my parents and my girlfriend for their love and support

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor, Dr Ross Maciejewski, who has been an invaluable mentor guiding me in my research. I could not have imagined having this thesis completed without his immense knowledge and expert suggestions. I would also like to thank my committee members, Dr. Huan Liu, Dr. Elizabeth Mack and Dr. Hasan Davulcu, who have provided me insightful comments and supported me further in my endeavors. Furthermore, I would like to mention that some of the material presented here was supported by the NSF under Grant No. 1350573 and in part by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001. I also would like to thank my colleagues at the Visual Analytics and Data Exploration Research (VADER) Lab for their constructive criticism and helpful suggestions regarding this work as well as for their support. Last but not least, I would like to thank my family and friends who have provided me with their love and affection and have believed in my abilities. They been a pillar of strength behind me through the years allowing me to focus and achieve my goals.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 INTRODUCTION 1

1.1 Motivation 2

1.2 Research Goals and Contributions 5

2 RELATED WORK 8

2.1 Spatiotemporal Visualization 8

2.2 Time Series Similarity 14

2.3 Map Classification 15

2.4 Clustering Evaluation and Comparison 17

2.5 Geographical Variation 18

2.6 Clustering with Feedback & Direct Manipulation 20

3 DOMAIN CHARACTERIZATION AND DESIGN CHOICES 22

4 VISUAL ANALYTICS FRAMEWORK FOR SPATIOTEMPORAL CLUSTERING 25

4.1 Data Wrangling 26

4.1.1 Temporal Configuration 26

4.1.2 Spatial Aggregation 29

4.2 Exploratory Data Analysis 32

4.2.1 Transformation and Normalization 33

4.2.2 Focus+Context Choropleth 33

4.2.3 Histogram 34

4.2.4 Box Plot 36

CHAPTER Page

4.2.5 Scatterplot 36
4.2.6 Level of Detail 37
4.3 Temporal Similarity 38
4.3.1 Temporal Trend Matching 40
4.3.2 Logic Tree 44
4.3.3 Temporal Trend Multiples 46
4.3.4 Comparison to Previous Work 47
4.4 Multivariate Similarity 48
4.4.1 Multidimensional Distance Map 49
4.4.2 Multivariate Clustering 50
4.5 Interactive Clustering 52
5 CLUSTERING EXPLORATION 57
5.1 Group Selection 58
5.2 Visual Exploration Widgets 59
5.3 Clustering Comparison 61
6 VISUAL APPEARANCE OF SPATIAL ASSOCIATION 67
6.1 Categorizing the Effects of Relabeling 69
6.2 Summarizing the Visual Impact 76
7 CASE STUDIES 78
7.1 Conflicts in West Africa 78
7.2 Crime Estimates in the United States 80
7.3 Indices of Industrial Diversity 83
7.4 Impact of Geographical Variations on clustering 87
7.4.1 Discrete Spatial Extent 88

CHAPTER Page

7.4.2 Discrete Geographical Features 90
7.4.3 Continuous Spatial Extent 91
7.4.4 Continuous Geographical Resolution 93
7.5 Visual Impact of Changes in Classification Boundaries 94
7.5.1 Applying EOC for Visual Clustering 94
7.5.2 Combining EOC and VIOC 98
7.5.3 Relabeling versus Boundary Modification 99
8 CONCLUSION AND FUTURE WORK 102
REFERENCES 107

LIST OF TABLES

Table Page

2.1 Overview of Relevant Analytical Toolkits 9

LIST OF FIGURES

Figure		Page
1.1	An Illustration for the Concept of Visual Analytics.	2
4.1	An Overview of the Framework Interface.	25
4.2	An Example of a Raw Spatiotemporal Data Table in .csv Format.	27
4.3	An Illustration of the Data Cube.	27
4.4	An Illustration of the Attributes Group Tree.	28
4.5	An Example of the Data Wrangling Process for Configuring Temporal Variables.	29
4.6	An Illustration of Two Spatial Aggregation Types in the Data Wrangling Process.	30
4.7	Demonstration of the Effects for Transformation.	32
4.8	An Example of the Focus+context Choropleth Map.	34
4.9	Illustration of Various Exploratory Data Analysis Widgets.	35
4.10	A Close Look at the Scatterplot in PCA Projection Mode.	37
4.11	An Example of Explorations at Different Levels of Details.	38
4.12	An Example of Time Series Similarity Computed with Euclidean Distance.	40
4.13	An Example of Time Series Similarity Computed with Sequential Normalized Euclidean Distance.	42
4.14	The Temporal Trend Definition Widget Allows Users to Interactively Customize Temporal Attributes Including the Temporal Value, Time Series Length, Lead and Lag.	43
4.15	The Similarity Logic Tree Widget.	45
4.16	Temporal Trend Multiples.	47

Figure Page

4.17 Here the User Is Exploring the Effects of Applying the Multivariate Similarity Tooltip. 50
4.18 Here Is an Example That Illustrates a Dendrogram and the Multivariate Clustering Results Using Hierarchical Clustering and k-means Clustering. 51
4.19 An Application of the DBI Value in the Framework after Running the Same Clustering Algorithm Multiple Times. 53
4.20 An Example of User-guided K-means. 56
5.1 An Example of the PCP Area Profiler and Rose Plot. 60
5.2 The Coherent Clustering Color Mapping with Both Maps Having Five Clusters. 61
5.3 Comparing Two Clustering Results for the Same Group of 15 Objects. . 64
5.4 An Example of the Triple-D View (Drag and Drop Clustering Difference View). 65
6.1 Four Spatial Cases and the Effects of Changing a Single Unit. 69
6.2 An Example of Adjacent Changeable Regions. 72
6.3 A K-means Classification of US Census Variables Illustrates Boundary Elements and Their Corresponding Cases From Figure 6.1. 75
7.1 Aggregation of Conflicts in West Africa. 78
7.2 An Example of Using Regular Expression with Aggregation Key. 79
7.3 Using the Threshold Widget as an Overview Mechanism for Exploratory Data Analysis. 81
7.4 Users Explore Temporal Similarity Through a Value Based Filter. 82

Figure Page

7.5 Users Explore Temporal Similarity for West Virginia Through a Rank Based Filter. 83
7.6 Exploration of the Indices of Industrial Diversity Dataset Using Analytical Brushing. 85
7.7 User Defined Trend Exploration. 86
7.8 An Example of Exploration Between Clusterings in Different Geographical Locations. 88
7.9 An Example of Partial Clustering for Units of Different Geographical Feature. 90
7.10 An Example of Scale Effect on Clustering Around Cook County. 92
7.11 An Example of Clustering Results under Different Geographical Resolutions. 93
7.12 Minimizing and Maximizing the EOC of Elements near the Classification Boundary Using Criminal Incident Reports in Chicago, IL. 95
7.13 Minimizing and Maximizing the EOC of Elements near the Classification Boundary Using Us Census Data from the Western United States. 97
7.14 Maximizing the EOC Based on the VIOC near the Classification Boundary Using Us Indices of Industrial Diversity from the Western United States. 98
7.15 The Effects of Model Manipulation on Choropleth Map Classification. . 100
8.1 An Example of Temporal Tree Map for 5 Clusters over 12 Years. 105

Chapter 1

INTRODUCTION

The current instrumentation of smart systems and cyber-enabled infrastructure is leading to a generation of large-scale, real-time datasets that have the potential to provide new information and insight into a broad range of spatial dynamics. In the past, spatial measurements were relegated to single snapshots in time with Decennial Population Censuses and other governmentally collected statistics providing the basis for spatial analysis. However, the current influx of data, from Twitter feeds to traffic patterns to home electric consumption, now provides real-time updates at fine scale spatial resolutions. Such emerging spatial big data has the potential to provide new understandings and spur innovation. For example, a 2011 McKinsey Global Institute report [1] estimated that savings of $600 Billion annually could be achieved by reducing idling and fuel consumption through smart phone navigation. Such an analysis could only be done by analyzing localized spatiotemporal traffic patterns. Clearly, such real-time data can generate unprecedented insights into spatiotemporal interactions and flows [2]. Unfortunately, the scale of the data is a double edged sword. While it is clear that such information will be able to provide new insights into the world, the generation of these insights is increasingly difficult.

Traditionally, visualization is one of the most important and commonly used methods of generating insight into large scale data. Particularly for spatial data, the translation of such data into a visual form allows users to quickly see patterns, explore summaries and relate domain knowledge about underlying geographical phenomena that would not be apparent in tabular form. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [4]. As illustrated in

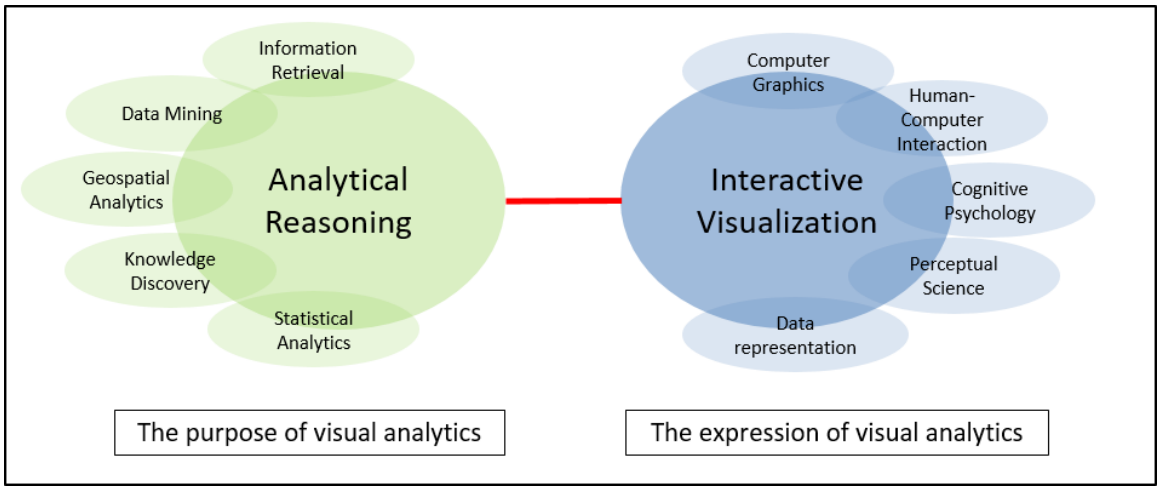


Figure 1.1: An illustration for the concept of visual analytics. This is similar to the concept figure in the work of Keim et al. [3].

Figure 1.1, visual analytics is a multidisciplinary field that consists of data mining, data management, human computer interaction and perception+cognition [5]. By utilizing human's cognitive and perceptual skills, visual analytics combines human factors into the decision making process and serves as a basis for communications between users and machines. Since visual analytics augments the analytical reasoning process upon the form of interactive visualizations (Figure 1.1), visual analytics tools enable decision makers to actually explore the nature of complex data and generate insight into large scale data. Due to the importance and complexity of spatiotemporal data, visual analytics approaches have been applied in a variety of spatiotemporal domain areas including healthcare [6–9], crime [10–12], environmental science [13–16] and socio-economic analysis [17–19] to study spatiotemporal processes and enable hypotheses generation .

1.1 Motivation

Several critical challenges arise when visualizing and exploring large spatiotemporal datasets. One of the major challenges is that spatiotemporal attribute data are

often organized very differently. They may be archived in a plain tabular format, such as .csv, or in a specific format, such as a .dbf file or a netCDF file, and all of those formats do not have a standard regulation on the arrangement among the attributes. This means that while analysts may know the temporal relationships among variables, machines may have no idea of the internal data relationship. Thus, designing a fully automated data generalization procedure becomes unrealistic. In order to apply visual analytics solutions, analysts have to put lots of effort into data preprocessing or manually coding scripts for data cleaning. Moreover, the fact that data are organized differently makes data fusion between different sources very costly. While many current geographic information systems (GIS) provide simple functions for integrating data based on specifying the joint geographical units, few data processing approaches handle the temporal aspect of spatiotemporal data.

As for visualizing spatiotemporal data, the underlying geographical component of the data lends itself well to univariate visualization in the form of traditional cartographic representations (e.g., choropleth, isopleth, dasymetric maps [20]). However, as the data becomes multivariate, cartographic representations become more complex. Multivariate color maps [21], textures [22], small multiples [23] and 3D views [24] have been employed as a means of increasing the amount of information that can be conveyed when plotting spatial data on a map. However, each of these methods have their own limitations. Multivariate color maps and textures result in cognitive overload where much time is spent trying to separate data elements in the visual channel [25]. Occlusion and clutter remain fundamental challenges for effective visual data understanding in 3D [26]. Utilizing small multiples can help in side-by-side comparisons but their scalability is limited by the available screen space and the cognitive overhead associated with pairwise comparisons [27]. Such problems are further compounded when a temporal aspect is added to the data. Such data is often

explored using animation and controlled playbacks; however, this forces end users to rely on their memory and do sequential comparisons between animation frames. As the number of spatial units being explored increases, small areal units can often be perceptually obscured within the data because size is often a dominating perceptual cue, thus make the data quickly become intractable. Furthermore, recent work has shown that map animation can lead to change blindness [28] when visually analyzing spatiotemporal data. Such challenges call for a transformative view to explore large spatial data. While intensive efforts have been spent on developing better algorithms and techniques for data management, querying and analysis of such data, much less attention has been paid to the design of effective visual analytics solutions.

Another challenge is that many data mining methods, such as clustering, have been introduced into the spatiotemporal visual analytics pipeline as an exploratory procedure, yet these methods are usually applied at a level that is agnostic to the spatial relationships between the data (i.e., the positions of the regions are not used as features in the clustering method). As such, local geographic variations may be obscured in a global clustering approach. Furthermore, the quality of the clustering results needs to be accounted for and explored both locally and globally. Choropleth maps using clustering approaches have been applied in various domain areas (e.g., GeoDa [21], VIS-STAMP [23]) for exploratory data analysis. Such maps allow multivariate clustering results to be interpreted based on the map appearance. One important aspect of the interpretation is the spatial association [29] related to Tobler's first law of geography [30] which states that "everything is related to everything else, but near things are more related." For instance, if the neighboring locations are mostly from the same cluster (in the same color), such an observation represents a strong spatial association; if the neighboring locations are scattered from different clusters (in different colors), such an observation represents more random-

ness in geographical space. In other words, the map appearance tells a story and the story could vary as the appearance changes. Knowing the stability of a choropleth map can be critical for interpreting the clustering result. By modifying the label of geographical elements, the resultant visualization can appear to have more (or less) spatial heterogeneity. This can lead to misinterpretation of the data and confuse findings and results [31]. However, due to the complex relationship between multivariate space and geographical space, there is still a gap between the multivariate clustering stability and the stability of the map appearance. While much work has been done on the evaluation of clustering stability and clustering tendency [32–35], little work addresses the visual stability when mapping the clustering result from a multivariate space to the fixed geographical space.

This thesis utilizes three different criteria and develops novel metrics to visually explore cluster labels near the cluster boundaries with a focus on the connection between the stability of the map appearance and the clustering. Instead of being confined to the original spatiotemporal domain, this concept extends traditional visual representations and presents novel views for showing how correlations, clusters and other various spatial dynamics change over time. In order to validate and explore more insight for those metrics, this thesis presents a novel visual analytics framework for spatiotemporal data exploration integrated with a data wrangling process.

1.2 Research Goals and Contributions

While intensive efforts have been spent on developing better algorithms and techniques for data management, querying and analysis of spatiotemporal data, much less attention has been paid to the design of effective visual analytics solutions. Instead of being confined to the original spatiotemporal domain, this thesis seeks to both extend traditional visual representations and develop novel views for showing how

correlations, clusters and other various spatial dynamics change over time. Underlying these novel views is the need for visual representations in which the manipulation of the representation is directly tied to the underlying computational analytics. The primary goals of this thesis include:

1. Enabling quick spatial data wrangling. As discussed in the motivation, being able to link datasets with varying spatial and spatiotemporal resolutions is challenging for a typical end user and tools that can address this need are critical.

2. Enabling improved pattern finding and hypothesis generation. From the visual analytics perspective, a pattern is the general term for any recognizable regularity in the data [36]. Regular structures that can describe changes in space and time, in particular, repeating structures, are often called spatiotemporal patterns [37].

3. Enabling cartographers and analysts to have a better understanding of multivariate map classification in geographical space and to create a more precise, accurate and meaningful map.

Furthermore, this thesis explores the design space for spatiotemporal analysis through the development of a comprehensive data wrangling and exploratory data analysis platform. Key contributions include:

1. A design study for system requirements of spatiotemporal analysis tools.

2. An implementation of a visual analytics framework for spatiotemporal data analysis with key features including a novel similarity query interface as well as a novel Drag & Drop clustering Difference view for cluster comparison.

3. A novel mathematical formulation for quantifying the visual impact of the classification boundary in choropleth maps.

4. A domain characterization for spatiotemporal exploration based on feedback from economic geographers and political scientists.

Chapter 2

RELATED WORK

This thesis presents a framework that is designed for the visual exploration of spatiotemporal data and the extraction of locally related spatial structures. There are several relevant fields to be reviewed including spatiotemporal visualization, time series similarity, map classification, clustering evaluation and comparison, geographical variation, clustering with feedback, and direct manipulation.

2.1 Spatiotemporal Visualization

Since the early 1990's, a number of approaches for dealing with spatiotemporal data have been proposed by the GISs (Geographic Information Science) community [38]. Table 2.1 provides an overview of some of these toolkits including their key features, their spatiotemporal capability and their respective programming languages. This table is not meant to be a complete review of all possible toolkits created since work in this area is quite extensive. Instead, this table is meant to provide an overview of toolkits that were foundational to the development of this thesis. A common feature of these toolkits is their one-directional integration with GIS [39]. One-directional integration means data are taken from a geographic information system and imported into a separate toolkit for analysis, or results are taken from a toolkit and placed back into a GIS for display and mapping [38].

Table 2.1: Overview of Relevant Analytical Toolkits

No.	Analytical Toolkit	First Citation	Related Citations	Year First Released	Updates	Example Application
1	REGARD (Radical Effective Graphical Analysis of Regional Data)	[40]	[41]	Not Released	N/A	Irish socioeconomic data (1991)
2	SPIDER (Spatial Interactive Data Explorer)	[42]	N/A	Not Released	N/A	Geochemical analyses (1991)
3	XGobi	[43]	[44, 45]	1996	Shifted to GGobi V2.0	Corn yield measurements (1997)
4	XmdvTool	[46]	[47]	1994	V8.0 released in 2010	Census data (2007)
5	SaTScan	[48]	[49, 50]	1997	V9.1.1 released in 2011	Epidemiological data (1997)
6	SAGE (Spatial Analysis in a GIS Environment)	[51]	[52, 53]	1997	No updates	Health data (1997)
7	CDV (Cartographic Data Visualizer)	[54]	[55]	Not Released	N/A	Educational attainment (1998)
8	CrimeStat	[56]	[57, 58]	1999	V3.3 released in 2010	Crime data (2006)
9	GeoVista Studio	[59]	[60, 61]	2002	V1.2 released in 2007	Forest habitat (2002)
10	SpDep (Spatial Dependence)	[62]	[63, 64]	2002	V0.5-53 released in 2012	African conflict data (2003)
11	GeoDa (Geographic Data Analysis)	[65]	[21]	2003	Shifted to OpenGeoDa V1.2.0	Homicide counts and rates (2006)
12	Dcluster	[66]	N/A	2004	V0.2-5 released in 2012	Sudden infant death syndrome (2003)
13	Improvise	[67]	[68]	2007	2011	Demographic data (2007)
14	Arc_Mat	[69]	[70]	2004	V1.0 in 2010	Population growth data (2004)
15	STARS (Space-Time Analysis of Regional Systems)	[71]	[72]	2004	V0.8.2 released in 2006	Regional income (2006)
16	VIS-STAMP (Visualization System for Space-Time and Multivariate Patterns)	[23]	N/A	2006	No updates	Company data (2006)
17	STAMP (Spatial-Temporal Analysis of Moving Polygons)	[73]	[74, 75]	2007	No updates	Wildfire spread (2007)
18	IVIID (Interactive Visualization tool for Indices of Industrial Diversity)	[76]	N/A	Not Released	N/A	Economy data

No.	Vis[1] (Y/N)	S[2] (Y/N)	T[3] (Y/N)	ST[4] (Y/N)	NS[5] (Y/N)	PL[6]	Data Type	Key Features	Link
1	Y	N	N	N	N	Pascal	Point, areal	Regional analysis, network analysis, animation, cross-layer linking and interactive graphics	http://www.statlab.uni-heidelberg.de/projects/workshop/Regardinfo.html

Continued on next page

No.	Vis[1] (Y/N)	S[2] (Y/N)	T[3] (Y/N)	ST[4] (Y/N)	NS[5] (Y/N)	PL[6]	Data Type	Key Features	Link
2	Y	Y	N	N	N	Pascal	Point	Multi-window, dynamic linking and ability of layers manipulation	No webpage link
3	Y	N	N	N	N	C	Point	Visualization engine, high-dimensional drawing, handle missing value, manipulation and display of the scatter plot	http://www2.research.att.com/areas/stat/xgobi/#xgobi-paper
4	Y	N	N	N	Y	C/C++	Non-Geographic	N-dimensional brushing, four methods for displaying multivariate data in both flat and hierarchical approach	http://davis.wpi.edu/xmdv/index.html
5	N	Y	Y	Y	N	Java	Point, areal	A flexible user interface for computing scan statistics for a variety of distributions to detect statistically significant clusters	http://www.satscan.org/
6	Y	Y	N	N	Y	C	Areal	Exploratory data analysis and exploratory spatial data analysis by utilizing GIS	ftp://ftp.shef.ac.uk/pub/uni/academic/D-H/g.old/sage/sagehtm/sage.htm
7	Y	Y	N	N	N	Tcl/Tk	Areal	Interative Cartographic visualization, comprises interpreted scripts for extension	http://www.spatial-modelling.info/CDV-Cartographic-Data
8	Y	Y	N	N	Y	C++	Point	Analyze the distribution, identify hot spots, indicate spatial autocorrelation, monitor the interaction of events and have specific crime analysis tools	http://www.icpsr.umich.edu/CrimeStat/about.html
9	Y	Y	N	N	Y	Java	Point, areal	Modular nature, codeless environment, combining computational clustering and sorting with cartographic and information visualization methods	http://www.geovistastudio.psu.edu/jsp/index.jsp
10	N	Y	N	N	N	R	Point, areal	A collection of various spatial analysis functions like regional aggregation, spatial autocorrelation, spatial regression model etc.	http://cran.r-project.org/web/packages/spdep/index.html
11	Y	Y	N	N	N	C++	Point, areal	Interactive environment that combines maps with statistical graphics and methods of descriptive spatial data analysis, such as spatial autocorrelation statistics, spatial regression	https://geodacenter.asu.edu/projects/opengeoda
12	N	Y	N	N	N	R	Areal	A set of functions for the detection of spatial clusters	http://www.uv.es/geeitema/Virgilio/Rpackages/DCluster/index.shtml
13	Y	N	N	N	N	Java	Areal	Declarative visual query language, multiple coordinated views and integrated metavisualization	http://www.cs.ou.edu/~weaver/improvise/index.html
14	Y	Y	N	N	Y	Matlab	Areal	Basic choropleth mapping and linked exploratory graphs combined with spatial data modeling	http://www.spatial-econometrics.com
15	Y	Y	Y	Y	N	Python	Areal	A number of recently developed methods of space-time analysis with an array of dynamically linked graphical views	http://regionalanalysislab.org/index.php/Main/STARS
16	Y	N	Y	N	Y	Java	Areal	Self-organizing map, combine visualization with clustering, sorting	http://www.spatialdatamining.org/software/visstamp

Continued on next page

No.	Vis[1] (Y/N)	S[2] (Y/N)	T[3] (Y/N)	ST[4] (Y/N)	NS[5] (Y/N)	PL[6]	Data Type	Key Features	Link
17	Y	N	N	Y	N	VB.Net	Point	Analyzing changes in multiple polygon layers inside ArcGIS, like phenomena that change spatially through time	http://www.geog.uvic.ca/spar/stamp/help/index.html
18	Y	N	Y	N	Y	C++	Areal	All interactive linked views, dynamic analytic filter, similarity computing and indices calculation	No webpage link

[1]Visualization [2]Spatial functionality [3]Temporal functionality
[4]Spatial-Temporal functionality [5]Non-spatial functionality [6]Programming Language

Some efforts have been made to link these packages to a GIS and/or statistical software to minimize the required transfer of data, but many of the toolkits in Table 2.1 remain stand alone packages. For example, Symanzik et al. [77] linked Arcview to the data visualization software XGobi, and they also expanded on this initial work to link the statistical software XploRe to Arcview and XGobi [44].

Other common features of the toolkits in the table are their interactivity via techniques such as linking and brushing for exploratory spatial data analysis (ESDA). For example, in GeoVISTA Studio selected counties on a map are also highlighted in a corresponding parallel coordinate plot (PCP) window [60]. In terms of ESDA, Crimestat is an example of a program that leverages classic EDA and ESDA tools for crime analysis. The hotspot functionality, in particular, uses several techniques including the local Moran's I [29] and k-means clustering [78, 79] to identify elevated areas of crime in an exploratory fashion.

A final item of note regarding Table 2.1 is the functionality of the interfaces for visualization, spatial, temporal, spatio-temporal analysis, and non-spatial analysis specified in the table as Vis, S, T, ST, and NS respectively. Vis corresponds to systems that are designed for visualization purposes and provide interactive graphics. These systems may or may not handle spatial or temporal attributes; however, their commonality is an interactive graphical display of data (as opposed to text only

reports). The Cartographic Data Visualizer (CDV) [54] is an example of this kind of toolkit. The abbreviation S corresponds to systems with an explicitly spatial component where spatial analysis is defined as techniques that account for spatial autocorrelation or analyze the underlying processes behind data with a locational component. This functionality includes variograms, Moran's I [80], Geary's C [81], the Getis and Ord G statistic [82], and the local Moran [29]. T refers to toolkits capable of analyzing temporal data via techniques such as similarity metrics, control charts, ARIMA modeling, and time series plots. ST in the table corresponds to systems that are able to import and analyze data with both spatial and temporal components. Although this definition does not correspond to true spatio-temporal analysis, where space and time are analyzed simultaneously, these systems represent great strides in overcoming the analytical challenges associated with spatio-temporal data mentioned previously. Of the toolkits highlighted in this table, just three were designed with original spatio-temporal functionality: STAMP [73], STARS [71], and SaTScan [48].

STAMP [73] is a toolkit that examines geometric changes for polygons where association through space and time is defined as the union between layers in consecutive time periods. Change is characterized as a series of events: generation, disappearance, expansion, or contraction and information about these events is stored as a field in a GIS layer. Global and local change metrics are also computed prior to the creation of the polygon change layer to characterize changes in polygons through space and time. STARS [71] is a Python-based toolkit for the spatio-temporal analysis of areal data and is comprised of two parts, a geocomputational module and a visualization module, which may be used together or separately. Key features of this toolkit include the Gini and Theil inequality measures as well as the capability of performing traditional and spatial Markov analyses.

Interestingly, both GeoDa and Crimestat were not originally space-time capable, but have been revised to include this functionality in later releases. Version 1.2 of OpenGeoDa is now capable of analyzing spatio-temporal data via map animation and comparative static box plots [83]. Users merely need to create variables in a .dbf for each time period of interest and then convert the table to a space-time project [83]. More current releases of Crimestat also make spatio-temporal analyses possible. Space-time tools within Crimestat include the STAC or the Spatial and Temporal Analysis of Crime routine [84], as well as the Knox and Bartlett [85] and the Mantel [86] tests for space-time interaction. Finally, the NS column defines capabilities of toolkits that analyze data but may not use the spatial component of data directly. Functionalities that support this type of analysis include k-means clustering, principal component analysis, and other data mining algorithms that do not explicitly look at the geographic space of the data.

Recently in other domains, such as ecotopes, Hargrove and Hoffman [13] explored clustering geographic regions based on their multivariate attributes using k-means to create geographic ecotopes. Maciejewski et al. [8] employed a similar approach to analyze multivariate healthcare trends. Rey et al. [87] developed Crime Analytics for Space-Time (CAST) program designated for crime analysis but useful for evaluating spatiotemporal trends and other variables of concurrent interest. Andrienko et al. [88] carried out research on creating meaningful and analyst-guided clusters of large collections of trajectories. They also developed a framework that focused on identifying geographic trajectories with similarity metrics [89] as well as a stack-based trajectory wall that utilized the space time cube for visualizing trajectories [90]. Pelekis et al. [91] designed a novel distance function as a similarity measurement for the analysis of movement data. More recently, Ferreira et al. [92] proposed model

that supports complex spatiotemporal queries over big spatiotemporal urban data (New York taxi trips).

There are many more examples of the visual analytics regarding the spatiotemporal data, thus several state of the art reports summarize the spatiotemporal visualization from the following different perspectives: forms of data representation [93], types of change in the data [94], both data and tasks [95, 96], visualization operators [97] or the process of the visualization [98]. Mack and Maciejewski [99] also summarized a profile of visualization toolkits regarding the broadband provision data.

2.2 Time Series Similarity

Overall, the goal of spatiotemporal visualization systems is to enable analysts to answer questions not only about where events and measures are occurring and their spatial relationships but also when. Since time moves in a linear fashion, geographical visualization systems often animate graphics to show the movement of trends over time. While animation provides an obvious way to display spatiotemporal data, it also introduces cognitive burdens [100] as the user now must retain information of the last state of the data visualization and compare it to the current state. Given these cognitive issues that can arise in map animation, this thesis explores the use of machine learning techniques coupled with interactive visualizations for identifying regions with temporal similarities.

Previous work in time series exploration has focused on utilizing similarity metrics for visualizing and discovering non-trivial patterns in large time series[101–103], and connecting time-oriented data and information to a coherent interactive visualization [104]. For example, Hochheiser and Shneiderman [105] presented a Timebox widget for specifying query constraints on time series data sets. Wongsuphasawat et al. [106] developed a new similarity search interface for temporal query specification,

and recent work by Alencar et al. [107] applied similarity based metrics and multi-dimensional scaling for time series exploration. In terms of temporal analysis in the geographic domain, Malik et al. [15] used Pearson's correlation coefficient to analyze temporal similarities in selected geographical regions; however, this technique only allowed for comparisons between two regions at a time. Hoeber et al. [108] developed GTdiff which utilized user defined binning and filtering to directly compare changes over time. Andrienko et al. [89, 109] proposed several methods focusing on the exploration of spatial distribution of temporal data to find similar local behaviors and pattern changes. However, their approach may not be able to capture bell shaped and wave like features. Thus, many methods that can find trends in time series have been developed such as Dynamic Time Warping (DTW) [110] and Edit Distance with Real Penalty (ERP) [111].

This thesis aims to improve on such results by enabling complex similarity searches over both space and time. This framework extends on previous work by allowing the user to specify locations and trajectories of interest through an interactive brush. It also supports user-defined temporal trajectory querying, lag and lead exploration, and interactive visual comparison such that users can explore complex questions. For example, "show me regions that have similar trajectories with regards to measurements A, B, C, and D". More details are described in Chapter 3.

2.3 Map Classification

Map classification is one of the most important aspects in spatiotemporal geovisualization. The goal of a classification scheme is to group similar observations and split dissimilar observations to simplify and clarify the message of the map [112]. For univariate data, the simplest methods include quantile, equal interval, and standard deviation [113]. More complex methods have been proposed since the early 1960's.

For example, Jenks developed natural breaks, which seeks to reduce the variance within classes and maximize the variance between classes [114]. Scripter presented nested means [115] that calculates intervals for statistical maps by repeatedly deriving and using the arithmetic mean to divide a numerical array. Cromley [116] proposed a minimum boundary error method that maximizes spatial similarity among contiguous units in the same class interval, and Armstrong [117] developed a genetic binning scheme that creates optimal classifications with respect to multiple criteria (e.g., number-line relationships, area covered by each class, fragmentation). The most important part of map classification is how to choose the breaks or class boundaries. Evans [118] categorized sixteen class-interval systems and suggested that class intervals should be selected according to the overall shape of the data distribution. Brewer et al. [6] compared seven map classification methods with fifty-six subjects in a two-part experiment to determine which classifications are most suitable for epidemiological rate maps. Sun et al. [119] proposed a heuristic classification approach that utilizes the class separability concept and other classification criteria. They compared their approach to other classification methods based on element separability; however, visual changes in the map appearance that could occur due to slight shifts in classification boundaries have not, to our knowledge, been fully addressed.

Multivariate map classification typically involves various data mining and machine learning classification methods (e.g., k-means and self-organizing maps). By assigning a color to each label/ class/ category in the clustering result, a choropleth map is generated. A well known example of multivariate map classification is demographics classification. Vickers and Rees created the United Kingdom National Statistics Output Area Classification (OAC) [120] which is an open geodemographic classification with a hierarchical structure of 7 super-groups, 21 groups and 52 subgroups. Scrucca [121] proposed a procedure for identifying spatial clusters by applying

k-means to a set of variables expressing local spatial autocorrelations. Recently, Andrienko et al. [122] developed a set of tools for visually analyzing map classification. Their tools allow users to specify arbitrary class boundaries with direct manipulation but do not provide support for automatically identifying elements on cluster boundaries. Other work from Andrienko et al. [123] showcased a framework based on a self-organizing map that can be analyzed in both spatial and temporal contexts, and Streit et al. [124] introduced a model-driven design process for the visual analysis of heterogeneous data.

This thesis explores how the classification boundaries impact the visual appearance of choropleth maps. Because of the efficiency and simplicity [125], k-means clustering remains the core algorithm for the computation of geodemographic classifications [126]. Therefore, we use k-means as our default multivariate classification method; however, our findings can easily be extended to other classification methods.

2.4 Clustering Evaluation and Comparison

Due to the non-intuitive association between multivariate space and geospace, clustering results (map classification) generated from high dimensional spatial data can be difficult to interpret and compare, for example, spatially grouping multivariate clusters together may lead to placing Alaska next to Texas or Great Britain next to Australia. This section reviews related work on evaluating and comparing clustering to better address such issues. For many clustering evaluation and comparison techniques, researchers assume a true cluster structure exists and use an external criteria of clustering quality, such as the Rand index [127] or NMI (Normalized Mutual Information) [128] to measure the concordance between the true structure and output of clustering algorithms [129–131]. Jung et al. defined clustering gain which is based on squared error sum as a measure for cluster optimality [132]. Their mea-

surements can be utilized to estimate the desired number of clusters for partitional clustering methods. Meilă [133] characterized some criteria for comparing two clustering results directly by treating clusters as elements of a lattice. However, those works still remain at the arithmetic level (i.e., only numerical indicators have been provided and no visual information is available for illustration). Hoffman and Hargrove [134] created a simple multivariate geographic clustering comparison according to their state space color assignments, yet they do not apply a uniform comparison method. Recently, Zhou et al. [135] extended parallel sets to provide the mutual comparison and evaluation of multiple partitions. Their visualization can present the overall change between clusterings but may not suitable for showing the detailed changes in geographical applications. Hu et al. [136] described a heuristic to promote dynamic cluster stability and maximize stability between labels. Their approach for visualizing multiple relationships ensures mental map preservation but lacks the capability to show detailed local comparison. Thus, to enable clustering comparison, this thesis introduces the novel Triple-D (Drag & Drop clustering Difference) View to interactively display visual results for cluster comparison.

2.5 Geographical Variation

One characteristic of multivariate clustering related with geographical data is that neighborhood relationships play an important role in the clustering outcomes. This thesis also focuses on exploring clustering under spatial constraints (e.g., neighborhood relationships), thus this section reviews the concepts and previous work on geographical variation and localized exploration. Data generating processes associated with spatial data are often characterized as spatial dependence or spatial heterogeneity [137]. *Spatial dependency* refers to the similarity in attribute values of nearby spatial units [137] as proposed in Tobler's first law of geography [30]. In contrast,

Spatial heterogeneity or *nonstationarity* refers to variation rather than similarity in values for a particular measures across all spatial units [138]. Spatial stationarity is often assumed in statistical analyses, but this is problematic in the presence of spatial heterogeneity where assumptions of a global trend do not reflect the underlying data generating processes [138]. As such, the diagnosis of local dependence and heterogeneity is particularly valuable to understanding statistical output.

Due to this persistent issue in spatial data, tools such as the Moran scatterplot [139], local indicators of spatial association [29], and geographically weighted regression (GWR) [140] are critical to diagnosing outliers which might otherwise be obscured in global and local statistics not designed to diagnose spatial heterogeneity. The development of the local Moran's I and GWR in particular were critical to analyses of spatial data because prior local statistics including the G statistic [82] and the G* statistic [141] are not capable of assessing spatial heterogeneity in the form of local outliers. Many other geographically weighted (GW) statistics have also been developed (e.g., GW summary statistics [142], GWPCA [143]). To visually diagnose local and global spatial dependence, Dykes and Brunsdon [144] introduced geographically weighted interactive graphics for exploring and hypothesizing the spatial relationships under different scale-based variations. Turkay et al. [145] have developed methods for exploring geographically referenced multivariate data over location and scale through a variety of linked small multiples and summary statistics. Goodwin et al. [146] developed a suite of novel interactive visualization methods to identify interdependencies in multivariate data coupled with a series of correlation matrix views. While Goodwin et al. focus primarily on spatial extents of pairwise correlations, this thesis explores spatial extents in the multivariate clustering space and enables exploratory analysis between clustering differences. Research has highlighted that multivariate results, when mapped, can produce non-sensical results [147] because closeness in multivari-

ate space is not necessarily the same as closeness in geographic space. Thus, the capability provided in the framework presented in this thesis enables the examination of spatial processes in clustering results and moves beyond potentially misleading visual inspection of maps and global summary statistics that obscure important local variations in multivariate data.

2.6 Clustering with Feedback & Direct Manipulation

In order to incorporate users' knowledge, such as pairwise constraints between class labels, more and more semi-supervised clustering approaches which can let users "guide" or "adjust" the clustering process have been developed. Unlike traditional clustering, the semi-supervised approach has a fairly short history [148]. Cohn et al. [149] presented an approach to incorporate user's feedback in the form of constraints used in future clustering iterations. Jain et al. [150] introduced a Bayesian feedback mechanism, Huang et al. [151] tackled four types of feedback in text clustering, and Balcan [152] combined a query-based clustering model to allow users to provide feedback in a natural way. Recently, Choo et al. [153] proposed weakly supervised nonnegative matrix factorization that can lead to semantically meaningful and accurate clustering results by taking various prior information into account.

Besides those model-side feedback solutions, interactive exploration of the clustering results combined with a visual analytics solution has become a major focus. Work in this area includes the VISTA system [154] which was developed to help domain experts validate and refine cluster structures through interactive feedbacks. VISTA allows users to mark the visual boundaries between clusters and refine the algorithmic result if applicable. Chen and Liu [155] developed iVIBRATE as an interactive machine learning tool which allows users to iteratively modify the clustering process with an adaptive labeling subsystem. Andrienko et al. [88] developed

methods for analyst-guided clustering of large collections of trajectories by combining clustering and classification together through an interactive interface. House et al. [156] developed a new framework that merged Bayesian statistics and visual analytics together called Bayesian Visual Analytics (BaVA) to foster learning from data and make cohesive visualizations adjustable. Leman et al. [157] extended BaVA to be a more general algorithm analogy referred as "Visual to Parametric Interaction" (V2PI), which can create data displays based on both mechanistic data summaries and expert judgment. Overall, the idea of V2PI is to enable users to make parametric changes to models that control visualizations while remaining in the visual data domain. Endert et al. [158] also explored two possible observation-level interactions within three statistical methods (e.g., Multidimensional scaling) such that users can express their reasoning on observations instead of on the model or parameters. Furthermore, Brown et al. [159] presented an interactive spatialization system specifically regarding the distance function. It allows experts to directly define a distance metric based on their understanding of similarity. Nevertheless, previous work rarely considers spatial feedback, such as neighborhood information, for multivariate clustering. This thesis focuses on the identification of class elements near boundaries and enable direct manipulation for relabeling class elements. While our focus is more on the resulting changes in the visual output (i.e., the choropleth map), the identification of elements that impact the visual output can be used as measures of importance to direct analysts' attention to elements that require further inspection.

Chapter 3

DOMAIN CHARACTERIZATION AND DESIGN CHOICES

One of the major contributions of this thesis is exploring domain needs for spatiotemporal analysis and then providing case studies to see how these needs can generalize to a variety of domains. The two critical domains that were explored using the developed framework were spatial econometrics and political geography. One major component identified in these domains is analytical brushing. In both cases, the data characteristics primarily focused on identifying regions with similar trends over multiple variables over space and time. In the case of economic geography, I have interviewed a single domain expert who was interested in finding regions of growth or decline in various sectors and exploring regional similarity in both spatially proximal and spatially non-proximal regions. In the case of political science, I have interviewed two domain experts who were interested in exploring armed conflicts and their relationship to underlying infrastructure.

The characterization of the needs from domain experts can be summarized as:

1. In general, experts would like to filter out specific areal units based on user-defined criteria as a first cut for generating descriptive statistics and general theory building. The primary focus is on discovering areas that have similar/dissimilar trends in terms of multivariable x, y and z.

2. Experts would like to know if particular regions have commonalities/discontinuities at a specific point in time that they believe is theoretically meaningful, especially for comparative politics.

3. As political violence and economic development have strong spatial contagion aspects to temporal lags and spatial lags, experts want to see how the lag of temporal trends affects other trends across geospace (i.e., the diffusion speed of temporal correlation in geospace).

4. Experts also wish to find the top N most similar regions based on a set of criteria. For example, show me the top five most similar regions to A in terms of variable x, y and z.

This domain feedback helped inform an iterative design process for framework development and user needs were directly incorporated into various framework features. Such collaborative work enables the building of tools that can be directly useful for domain experts while providing a platform for basic research exploration. Here are several key design choices based on the domain needs.

- *Choice of queries over animation.* The first domain need is to identify similar trends for different variables. When tracking a limited number of temporal trends, users have to focusing on select areal units in the data [160] and mentally combine information across views or animation in order to determine trends and patterns (e.g., GeoDa). Based on interviews with our domain experts, they noted frustration in trying to track and compare multiple regions on the map when data was animated. As such, our analytical brushing framework was designed to enable them to define search criteria with respect to multivariate spatiotemporal trends. Similarity queries are bound to the mouse tip in order to enable focus+context exploration and improve visual clarity. In this manner, users can refine their queries based on a variety of operations (greater than, less than, AND, OR, XOR). This method can maintain the trace of the filtering operations to enable quick transitions between subsets and supersets.

- *Choice of user-defined similarity threshold over clustering.* Upon exploration with the analytical brushes, our domain experts indicated a need for more "fuzziness" in the similarity criteria, as well as uncertainty as to why certain areas were seen as similar. Such concerns about the black-box nature of many machine learning algorithms underscored the need for a way to insert the human into the similarity guidance loop. This suggests the need for a visual comparison of similarity results so that users can quickly understand what is meant by "80% similar" and whether their intuition about similarity is being met by the algorithmic criteria. This led to the implementation and use of the small multiples trend graph. In this manner, analysts can empirically inspect the similarity distribution and define their own similarity threshold. In this way, our framework can support the second domain need of finding similar/dissimilar trends.

- *Choice of distance map and small multiples.* While our experts indicated that the small multiples graph helped in understanding what was meant by similarity, they also indicated that they would like to better tie this back to the geographical exploration. As such, the distance map was developed to represent the temporal similarity distance to all other geographical areas from the brushed unit. In this way, users can determine the spatial compactness of a region with regards to a temporal distance function. While the distance map focuses on the overall spatial distribution, the small multiples focus on spatial distribution of local regions with more detail on the temporal representation. In this way, experts can explore the spatial distribution of similarity in our framework.

Chapter 4

VISUAL ANALYTICS FRAMEWORK FOR SPATIOTEMPORAL CLUSTERING

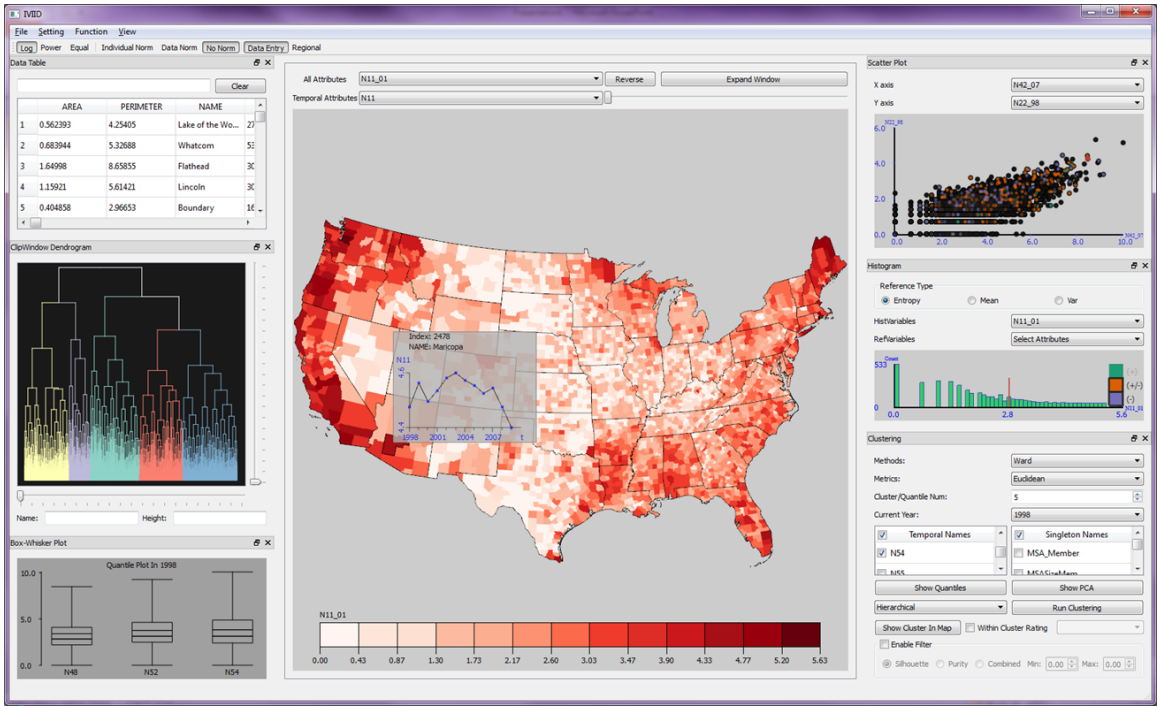


Figure 4.1: An overview of the framework interface. This framework provides an interactive scatterplot, histogram, tabular, quantile plot, dendrogram and geographical viewing widgets.

Figure 4.1 illustrates our visual analytics framework. This framework provides interactive scatterplot, histogram, tabular, quantile plot, dendrogram and geographical viewing widgets. Linked brushing [161, 162], dynamic graph tick-marks [163], Color Brewer color schemes [164] and various focus+context highlighting methods [165] are integrated into the framework. All widgets are dock-able and resize-able to allow for customized analysis. To enhance the multivariate data analysis process, our frame-

work also supports various clustering methods (e.g., hierarchical clustering, k-means Clustering) along with several types of distance metrics (e.g., Euclidean, Manhattan). In this chapter, I will introduce the main features of the framework.

4.1 Data Wrangling

Data wrangling is the process of restructuring raw data into a proper format such that the data will be palatable to certain tools. From the research report of Kandel et al. [166], data wrangling is the most tedious part for data scientists during their data analytic process and wrangling is responsible for up to 80% of the development time and cost. Therefore, Kandel et al. have developed Wrangler, an interactive system for creating data transformations [167]. Their system leverages semantic data types to aid validation and combines direct manipulation of visualized data with automatic inference of relevant transformations. However, their tool is not specified for spatiotemporal data. Spatiotemporal data not only have high dimensionality but also contain complex relationships between space and time. This thesis implements an open wizard as an effective method for the wrangling process of spatiotemporal data including components for temporal configuration and spatial aggregation.

4.1.1 Temporal Configuration

Usually raw spatiotemporal data does not explicitly indicate temporal information in their data structure. For example, data records often are in the form of a plain .csv file with some of the attributes' names bound to a timestamp (highlighted text in Figure 4.2, those numbers stand for year). Human beings can often distinguish such temporal information, yet machines must be programmed to understand such internal variable relationships, e.g., the granularity of time (are those numbers years,

AREA	PERIMETER	NAME	FIPS2	MSACOD_(	MSA	MSASize	Urb_	TEST98	TEST99	TEST00	TEST01	TEST02	TEST03	TEST04	TEST05	TEST06	TEST07
0.56239327872	4.25405095316	Lake of the Woods	27077		0		0	157	159	157	149	155	153	151	151	158	157
0.68394438598	5.32688084136	Whatcom	53073	13380	1	S	1	5317	5386	5386	5423	5583	5680	5950	6119	6313	6519
1.64997868083	8.65854773239	Flathead	30029		0		1	3051	3081	3	3279	3410	3594	3774	4041	4223	4364
1.15921172974	5.61421195338	Lincoln	30053		0		1	577	581	5	559	582	576	606	629	635	683
0.40485818528	2.96652641948	Boundary	16021		0		1	354	358		336	342	333	355	361	383	412
1.33426147566	5.51799347690	Blai														148	146
0.79916713373	5.26466994772	Stev														1024	1012
1.67672358996	7.11364215968	Oka														1198	1195
0.35560480739	2.58827605551	Pembina	38067		0		0	329	339	333	315	326	324	320	317	312	311
0.34940041914	2.69060467344	Kittson	27069		0		0	171	162	160	154	163	160	158	159	156	151
0.92141502553	4.56565369667	Hill	30041		0		1	524	518	503	505	517	511	516	519	521	534
0.47856051012	3.05848042013	Cavalier	38019		0		0	183	181	183	186	171	174	173	172	163	166
0.71103268191	4.15322218034	Ferry	53019		0		0	149	158	149	143	142	137	144	132	134	142
0.44948868499	3.09588721028	Pend Oreille	53051		0		0	232	233	219	237	258	251	252	249	265	263
1.63508313902	6.88142680936	Phillips	30071		0		0	149	144	151	142	146	149	151	152	145	146
0.32946487744	2.36861702152	Towner	38095		0		0	107	96	94	94	95	94	97	99	92	86
0.53178678484	3.60659184326	Roseau	27135		0		1	391	393	397	389	402	414	427	422	422	418
1.59101124051	6.45018874180	Valley	30105		0		1	259	258	257	251	267	271	277	262	271	279
0.54010425163	3.24068747346	Sheridan	30091		0		0	164	159	147	151	146	140	145	141	140	146
0.41043840447	3.05824537843	Divide	38023		0		0	83	85	84	82	78	82	90	88	83	88
0.45208723981	3.22945959031	Daniels	30019		0		0	82	80	81	85	75	79	77	76	71	72

TEST98 TEST99 TEST00 TEST01 TEST02 TEST03 TEST04

Figure 4.2: An example of a raw spatiotemporal data table in .csv format. Here the red rectangle highlights the temporal variables that can be manually distinguished. The last two digits in the variable name are the last two digits for the year.

months or days?), the position of temporal variables (which of those variables contain temporal information?) and the order of temporal variables.

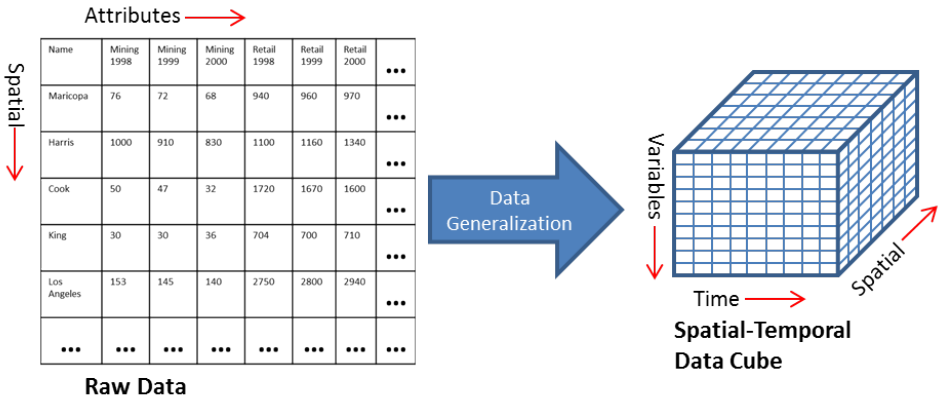


Figure 4.3: An illustration of the data cube. After the data wrangling procedure, data in plain format can be organized into a data cube and then stored in the framework.

In order to apply temporal operations, such as de-trending and similarity comparison, users need to be able to distinguish temporal variables from ordinary variables in a dataset and programmatically define such information. This thesis implements

a data wrangling procedure to transform tabular data into a uniform data cube for the internal framework data structure (Figure 4.3).

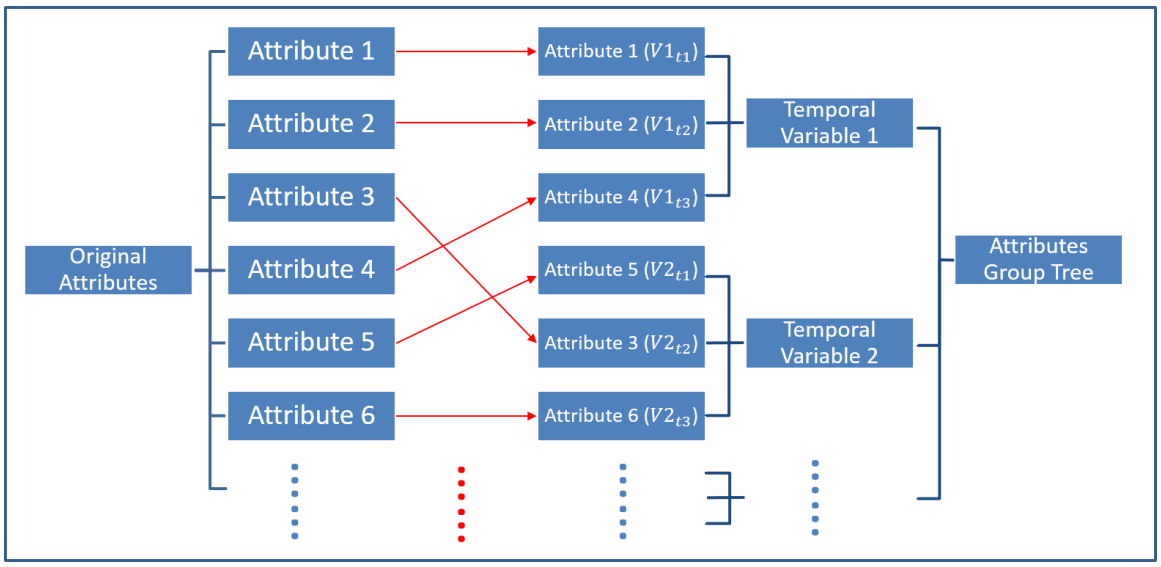


Figure 4.4: An illustration of the attributes group tree. Here $V1_{t1}$ stands for the first timestamp of the temporal variable 1, so on and so forth.

One of the main steps is configuring the temporal relationship via the attributes group tree (Figure 4.4). In the tree view, the original attributes are listed in the same level. By selecting and grouping the homogeneous attributes, the grouped attributes will be formed together as a new temporal variable for future use. The temporal sequence is determined by the order of positions in the temporal variable. For example, in Figure 4.5(1), a user loaded a .shp file and its corresponding .dbf file into the system. Even though the .dbf file contains temporal data which are explicitly noted in the attributes names (Figure 4.5(2)), the system would not recognize the correct temporal structure without human intervention. This framework allows users to inspect the data attributes and group them into temporal variables. To facilitate the wrangling process, this thesis adds a regular expression field where users only need to specify the common names for the temporal variables and the framework

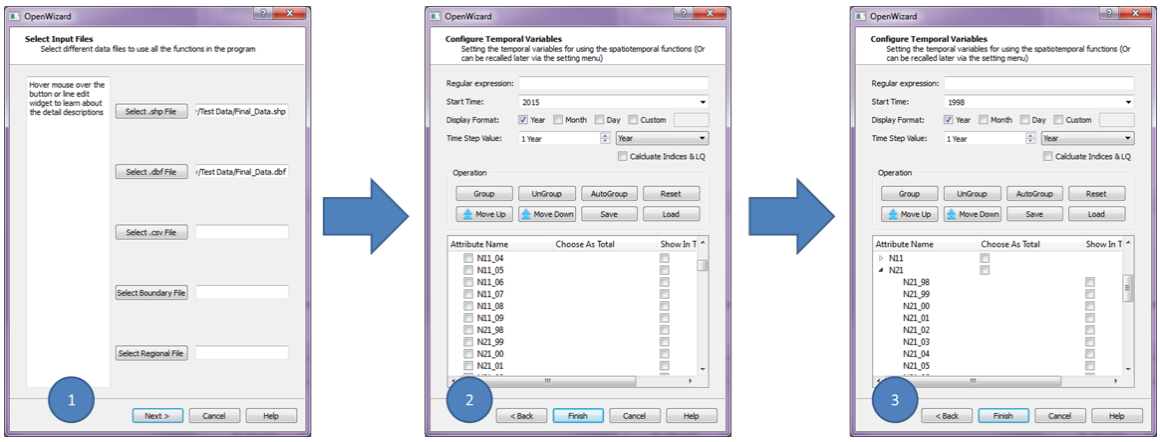


Figure 4.5: An example of the data wrangling process for configuring temporal variables. Here it shows the detailed procedure from left to right. The left figure indicates a page which allows for file selections of different formats. The middle figure shows a page that loads in all the plain text attributes and allows for grouping operations. The right figure shows an example of the result page after the grouping operations.

will group variables based on that expression automatically. Other related temporal properties (e.g., the start time, the time step interval) can be edited through the configuration part at the top of the open wizard as in Figure 4.5(3).

4.1.2 Spatial Aggregation

Raw spatiotemporal data may not come from a single file. It could be collected in multiple files of different format. Thus, this framework integrates spatial aggregation into the data wrangling procedure for combining data from different sources. The spatial aggregation process in this framework allows users to aggregate on different levels of spatial objects and generate user-defined variables on the fly. There are actually two types of spatial aggregation in this framework:

1. *Aggregation based on coordinates:* This type of aggregation requires the incoming data to have geo-coordinate attributes. Usually this is good for point type

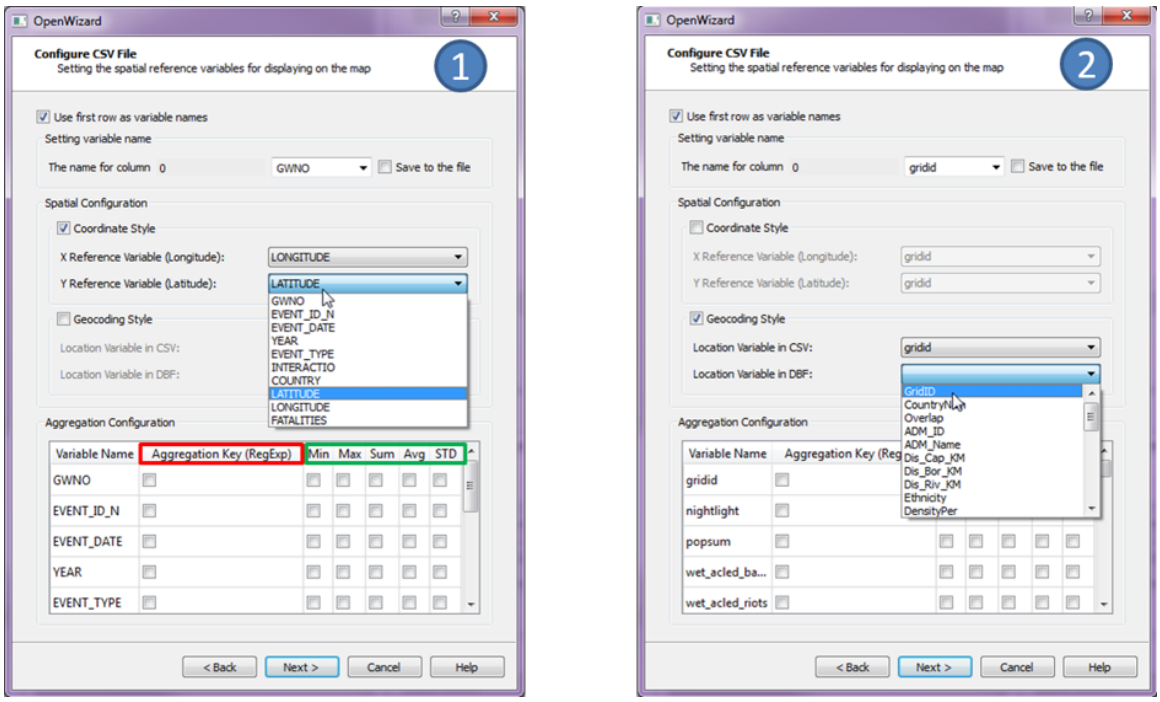


Figure 4.6: An illustration of two spatial aggregation types in the data wrangling process. The left figure shows the aggregation based on coordinates. The right figure shows the aggregation based on geocodes.

data such as incidents or crime events. Users only need to specify the corresponding data attributes that represent geo-coordinates before the aggregation process as shown in Figure 4.6(1).

2. *Aggregation based on geocodes:* This type of aggregation requires the incoming data to have consistent identification fields for the base data which could be synthetic ID numbers or geocoded names (e.g. state names, country names). This type of aggregation is usually suitable for areal type data. Users only have to specify the corresponding location variables before the aggregation process as in Figure 4.6(2).

Both types of aggregation allow users to define the operations for each of the attributes in the incoming data. There are six pre-defined aggregation operators (highlighted by a green rectangle in Figure 4.6(1)): count, average, minimum, maximum, sum, and standard derivation. When an aggregation operator has been checked, the corresponding variable will be derived using that operator. For example, when aggregating a crime events data file to the United States, the crime data contains 5 variables: case number, longitude, latitude, crime type and deaths. By choosing the count as the aggregation operator for the case number, users can know how many crime events happened in certain states. By choosing the average as the aggregation operator for the deaths, users will learn about the average death in a certain state. Often times, users would like to perform higher levels of aggregation that can be customized by certain data attributes, such as types of crime. This framework allows users to specify the aggregation key in the process (highlighted with red rectangle in Figure 4.6(1)). The aggregation key will be automatically scanned and stored in a hash table and only the entries with the same key will be aggregated. In the previous example, users would also like to know the average deaths by each crime type. Then they can check the aggregation key for the crime type attribute and select "Avg" as the aggregate operator to derive the average deaths by each crime type for each county. Eventually, the newly derived variables from the aggregation process will be renamed in the format of VN_AK_AO and combined with the original data. Here VN is the placeholder for the aggregated variable names, AK is the placeholder for the aggregation key if any, and AO is the placeholder for the aggregation operator. Moreover, this framework provides the capability of matching regular expression in the aggregation key field. This means instead of the whole attribute value/text, only the captured value/text by the regular expression will be stored in the hash table and used as an aggregation key.

4.2 Exploratory Data Analysis

By combining various visual representations and analysis tools, users can begin generating insight into how or why something is occurring and begin observing emergent patterns. Traditional EDA (Exploratory Data Analysis) [168, 169] techniques include line charts, box plots, histograms, and scatterplots. However, these techniques not only need additional modifications to adapt to the spatiotemporal context, they also face the same problems of cognitive overload when data becomes large. This framework extends those methods in different aspects to improve the visual efficiency.

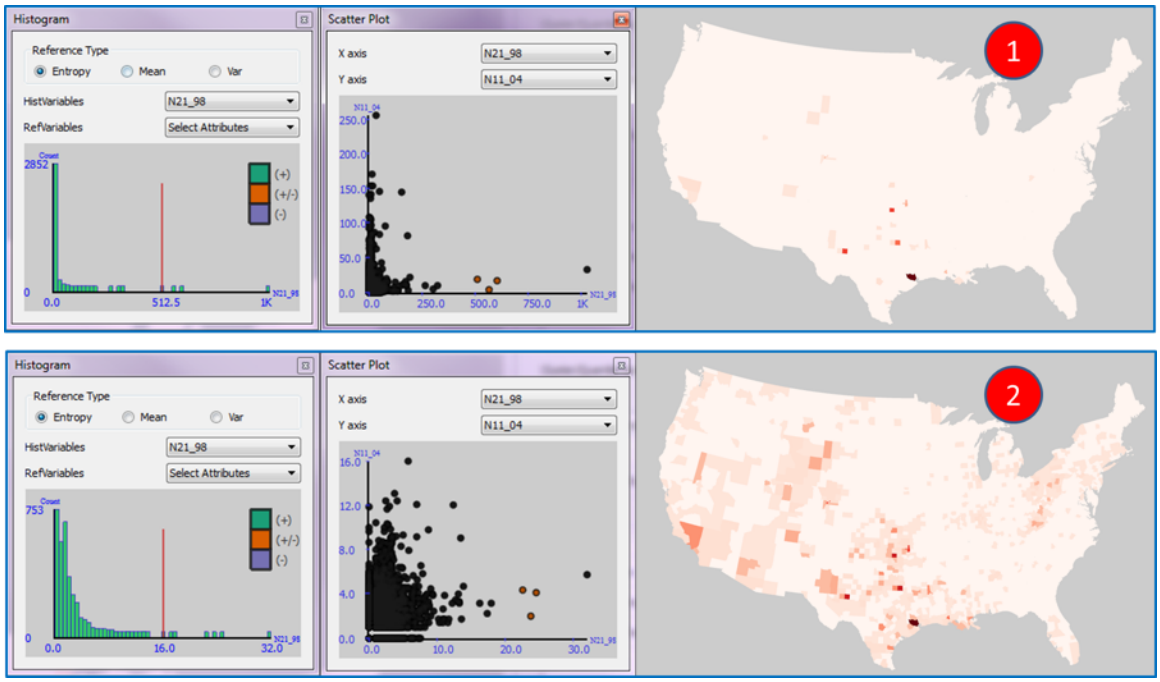


Figure 4.7: Demonstration of the effects for transformation. The top figure shows the histogram, scatterplot and choropleth map for the raw data without any transformation. The bottom figure displays the same widgets for the raw data after transformation.

4.2.1 Transformation and Normalization

One major problem during exploratory data analysis is that the data distribution is often skewed. Kasik et al. [170] stated that in order to allow users to obtain insight from visual representations, a wide range of algorithmic approaches are needed to transform the raw data into increasingly concise representations. For example, Figure 4.7(1) shows the histogram, scatter plot and choropleth map of the original data which are seriously skewed. In contrast, Figure 4.7(2) shows a different story with the same visualization techniques for the same data but uses a power transformation. In Figure 4.7(2), the choropleth map reveals a much more dynamic spatial distribution. Furthermore, normalization is required by many data mining techniques for improving the performance and effectiveness [171]. In order to enhance the visual efficiency and data mining performance, this framework provides global data transformation and normalization on the fly. Whenever a transformation or normalization is changed all the linked widgets are updated as well.

4.2.2 Focus+Context Choropleth

The choropleth map is a powerful technique to visualize how a measurement varies across a geographic area. However, when the number of spatial units becomes large, users may have difficulties when trying to identify regions of interest. As such, combining focus+context techniques with choropleth maps is necessary. The idea of focus+context is to allow users to have both overview and detail information simultaneously [172]. By preserving the global view at reduced detail (context) and maintaining the selected regions in greater detail (focus), focus+context helps users better understand how the points of interest relates to the entire data structure. This framework implements two focus+context techniques which are the *floating thumb-*

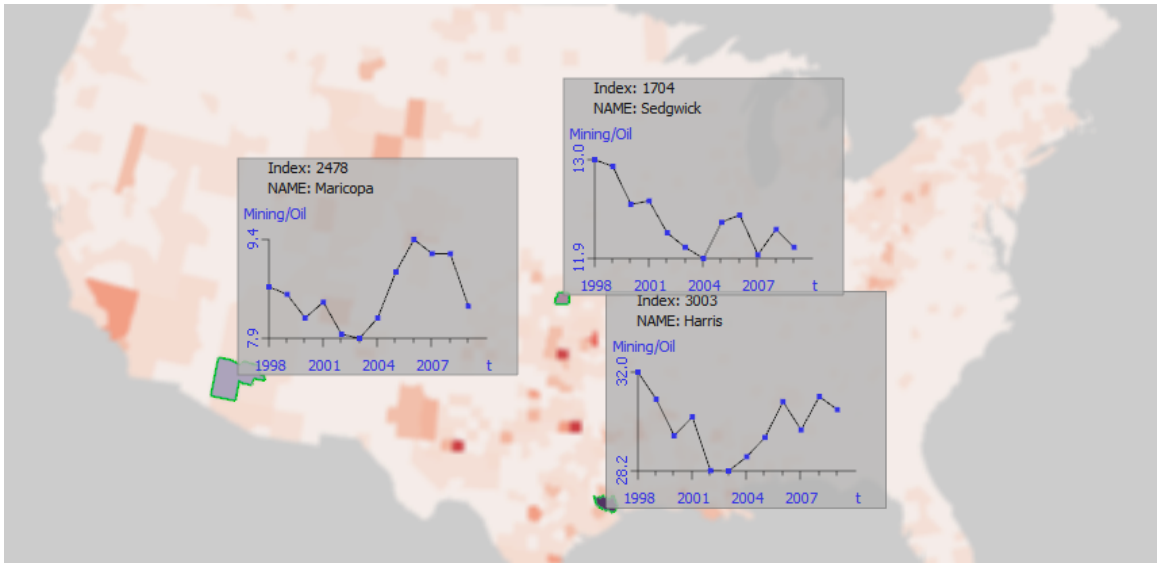


Figure 4.8: An example of the focus+context choropleth map. The thumbnail plots are displayed for the selected regions as the focus, while other regions get blurred out for preserving the context.

nail and the *blur highlighting*, for improved visual efficiency. When users mouse over a spatial unit, the unit's floating thumbnail will pop up and display linked important properties of the spatial unit chosen by users, such as temporal trend lines. The floating thumbnail is semi-transparent and can be enabled or disabled as well as fixed at certain positions. Figure 4.8 shows three floating thumbnails that are fixed on their three selected counties respectively. This allows users to quickly inspect the temporal trends of interest or perform analytical brushing. Blurring is also an effective way of highlighting [165], and this framework combines blurring with border coloring to enable users to quickly identify points of interest.

4.2.3 Histogram

A histogram is a graphical representation of the data distribution. This framework extends the histogram by adding a temporal threshold which allows users to interac-

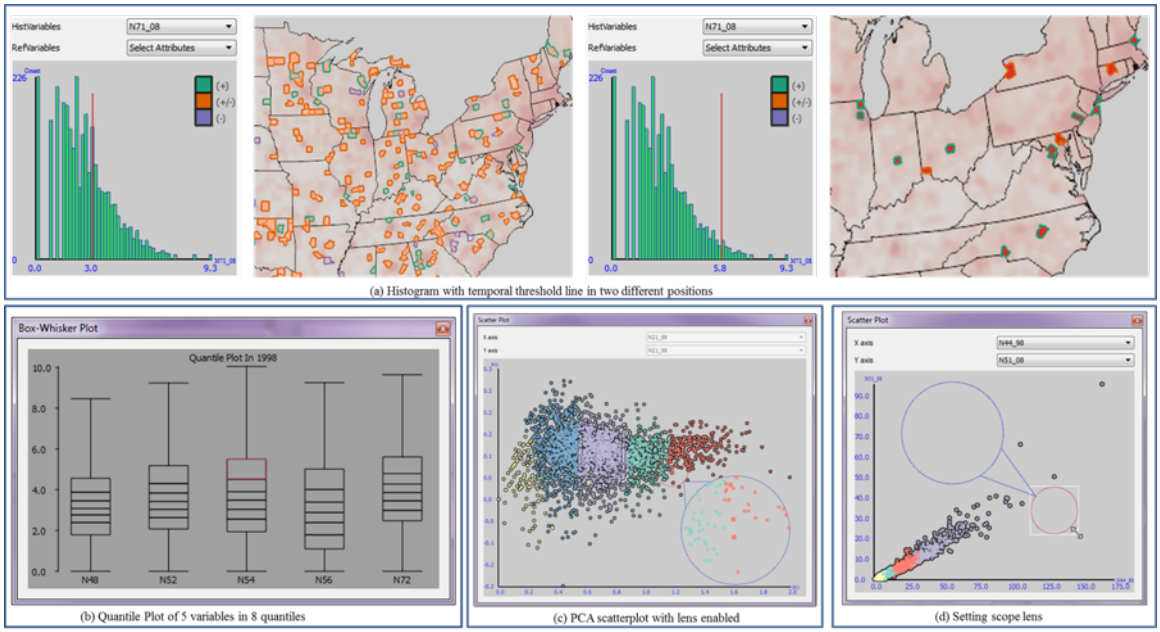


Figure 4.9: Illustration of various exploratory data analysis widgets. The top row shows the histograms with temporal threshold at different positions. The bottom figures show other EDA widgets including the box plot and scatterplot.

tively search for changes within a temporal variable by manipulating the vertical line in the histogram view illustrated in Figure 4.9 (a). If the user wishes to determine if an area has an upward, downward or oscillating trajectory, the user can interactively adjust the position of the vertical line. Based on the position of the line, spatial areas are then assigned to three classes. The first class (colored green and labeled with the + sign) indicates that at some point in time the area's statistical measure will cross from below the threshold set to above the threshold, indicating a positive trajectory. The second class (colored purple and labeled with the - sign) indicates that at some point in time the area's statistical measure will cross from above the threshold to below the threshold, indicating a negative trajectory. The third class (colored orange and labeled with the +/- sign) indicates that an area's statistical measure will oscillate around this value, crossing this threshold multiple times. By adjusting the

threshold value, the user can search for spatial clusters or disparate related areas. All spatial units that cross this threshold are outlined with the corresponding colors, and all other spatial units are blurred out in a focus+context manner. This method is useful for getting an overview of temporal trends with categories if users have a specific threshold or critical value such as the poverty line. The system allows those three classes to be enabled or disabled at any time to reduce the clutter in the visual representation.

4.2.4 Box Plot

The box plot is a convenient way of graphically depicting groups of numerical data through their quartiles. However, quartiles are usually not able to show the distribution of the data in as much detail as quantiles. Thus this framework extends the box plot to be an interactive quantile plot. This allows users to set up the proper quantile number and inspect any of those quantiles in their favor (Figure 4.9 (b)). By double clicking in the corresponding quantile, all the spatial units in that quantile will be highlighted on the map. This widget applies to multiple variables by lining them up horizontally, and allows users to compare the distribution of the attributes from the same temporal variable.

4.2.5 Scatterplot

Scatterplots are usually designed for bivariate data. In this framework, the scatterplot is enhanced with two modes: one is the bivariate mode while the other is the projection mode that utilizes Principle Component Analysis (PCA [173]). PCA also provides helpful information when users try to identify border points in clustering. For instance, Figure 4.10 shows a k-means clustering of 4 variables. From the PCA scatterplot, users can quickly tell where the point is and which border the point it is

lying on. The border points are highlighted simultaneously in the map with a black outline and in the PCA plot with a rectangular shape.

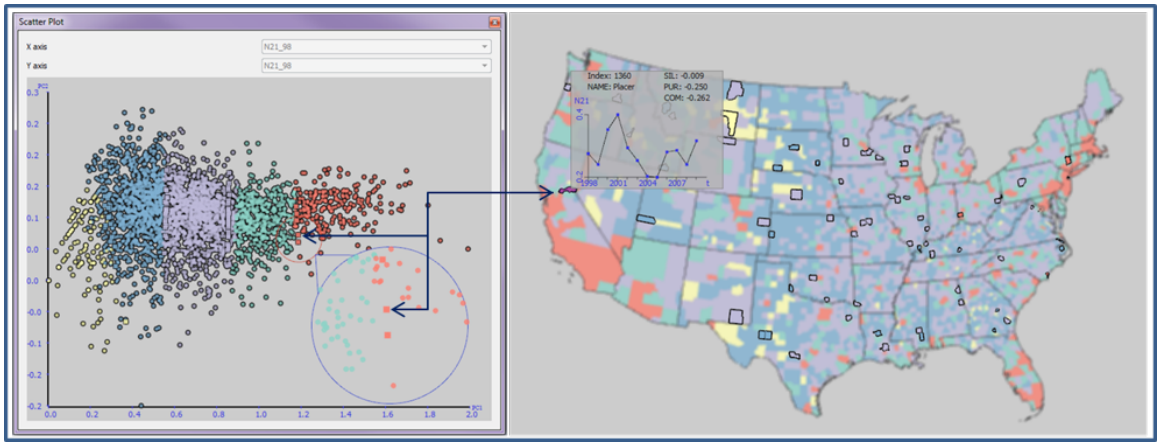


Figure 4.10: A close look at the scatterplot in PCA projection mode. Placer County gets selected in the choropleth map, the scatterplot, and the adjustable scope lens.

In addition to the PCA scatterplot, this framework also uses an adjustable scope lens (Figure 4.9 (c)) for tackling the clutter problem when spatial units dramatically increase. The lens consists of a configuration part and an application part. Both parts can be freely panned and scaled in the setting mode (Figure 4.9 (d)). By setting the ratio between the zoom area and the lens area, users can create a detailed view of the scatterplot in any zoom level without modifying the original plot resolution.

4.2.6 Level of Detail

Spatial resolution is one important characteristic of spatiotemporal data. To enable users to flexibly explore spatiotemporal data under different spatial resolutions, the idea of LOD (level of detail) has been integrated into the EDA process. This feature is similar to spatial aggregation in the data wrangling procedure. However, instead of only generating new variables for the base spatial units, this process will also generate a new spatial layer. After that, all the operations and analytic methods

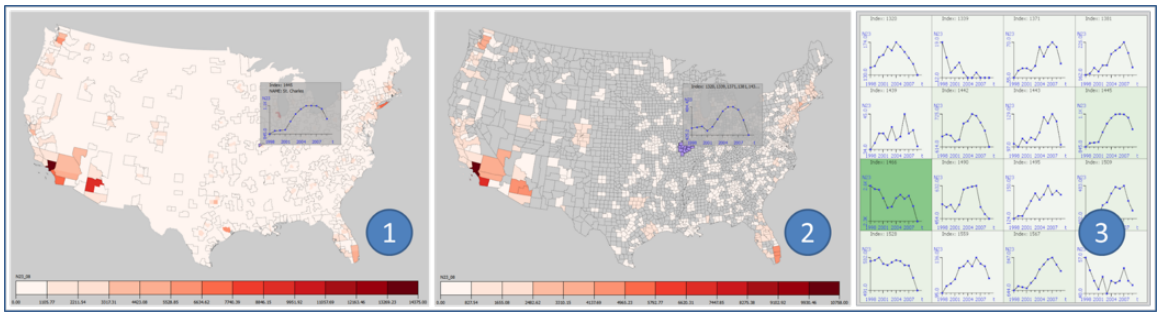


Figure 4.11: An example of explorations at different levels of details. The left figure shows the choropleth map for county level. The middle figure shows the map for metropolitan level. The right figure shows the thumbnail matrix for all the counties in the selected metropolitan area. Notice the difference between the temporal trend in the thumbnail of (1) and (2).

will be applied on the new layer. Users can switch between the new layer and the old spatial layer. LOD is very useful when users have high resolution spatiotemporal data but want to analyze the data or compare to other data at lower resolutions to gain an overview of the statistics. For example, the original data in Figure 4.11(1) are US counties. After applying the LOD, users can explore metropolitan areas which are made up of several counties as shown in Figure 4.11(2). Figure 4.11(3) displays a thumbnail matrix for all the counties in the highlighted metropolitan area of Figure 4.11(2).

4.3 Temporal Similarity

As the number of geographic units in the dataset increases, various problems emerge that hinder the traditional visual analysis of data. In choropleth maps, small areal units can often be perceptually obscured due to the color of neighboring regions. This problem is further exacerbated when exploring spatiotemporal data. The most common way to explore such data is by creating animated maps; however, animation

also introduces cognitive burdens [100] as the user must retain information from the last state of the data visualization and compare it to the current state. Recent work has also illustrated that map animation can lead to change blindness [28] when visually analyzing the data, and other work demonstrates that users are only able to track a limited number of temporal trends, focusing only on select areal units in the data [160]. Furthermore, the choice of class intervals becomes increasingly challenging when creating the animation of choropleth maps as it can potentially emphasize relatively small fluctuations due to temporally global choices in class selection [174]. In order to reduce these types of cognitive burden, many systems allow for either looped playback animation, or user controlled exploration through an interactive time slider. However, such exploration places the burden of discovery on the analyst.

This thesis proposes the use of brushing techniques that link multivariate and time series similarity metrics as a means of augmenting the traditional geographical visual analysis process. Here, a similarity query is tied directly to a brush tooltip so that users can interactively define the query and then explore similarity metrics via mouse over. In this way, users are able to direct their focus to locations that are known to have similar multivariate properties or similar temporal trajectories. Furthermore, to extend this to multivariate trends, this framework also uses logical operations in order to dynamically restrict comparisons across spatial units of the data. Thus, users are able to interactively answer traditional questions (such as, "show me the locations that have high/low values") through brushing and highlighting, as well as analytically query the data to answer more complex questions (such as, "show me the regions that have a temporal trend similar to region A" or "show me the areas that are similar in terms of multivariable x, y and z"). Such linked brushing is directly aligned with Keim's visual analytics mantra [3], "Analyze first, show the important, zoom, filter and analyze further, details on demand."

4.3.1 *Temporal Trend Matching*

Part of this thesis focuses on time series similarity as the primary means of searching for similar (or dissimilar) trends between different spatial regions. For a given region and a given temporal variable, a time series $\mathbf{p}$ can be denoted as $(p_1, p_2, p_3, \ldots, p_n)$. For any other region in the dataset, there exists another time series $\mathbf{q}$ denoted as $(q_1, q_2, q_3, \ldots, q_n)$. To identify asynchronous changes or patterns over time, settings for lag adjustment as well as pattern length are provided. The lag is a shifting operator between two time series, and a user may refine the beginning and end of the time range under analysis. Then, when the lag l is positive, $\mathbf{q}$ will shift backwards such that $\mathbf{p}$ is represented by $\{p_{1+l}, p_{2+l}, \ldots, p_n\}$ and $\mathbf{q}$ is represented by $\{q_1, q_2, \ldots, q_{n-l}, \}$; when the lag l is negative, the time series shifts forwards so the distance will only consider $\{p_1, p_2, \ldots, p_{n+l}\}$ from $\mathbf{p}$ and $\{q_{1-l}, q_{2-l}, \ldots, q_n\}$ from $\mathbf{q}$. In order to determine if other regions follow a similar temporal pattern as the region of interest, this framework utilizes time series similarity metrics from the data mining community [175].

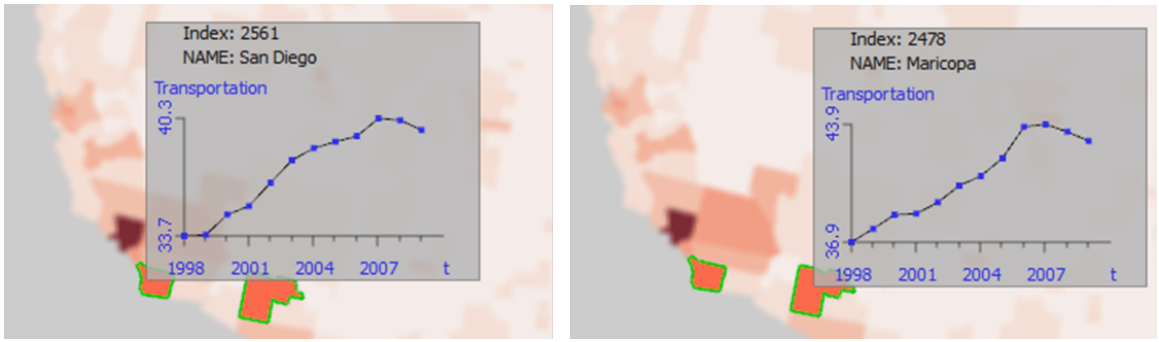


Figure 4.12: An example of time series similarity computed with Euclidean Distance. One can tell that the trajectory still have dissimilarity, but the range of temporal values are similar.

To facilitate such exploration, the system incorporates a variety of distance metrics with which the user can compute the time series similarity. The need for these different similarity functions is that each metric can provide the user with unique insight into their data. The chosen metrics include Euclidean Distance, which allows analysts to search for matches between temporal signatures and is defined as:

$$dist_{Euc}\left(\mathbf{p},\mathbf{q}\right)=\sqrt{\sum_{i=1}^{n}\left(q_i-p_i\right)^2}. \tag{4.1}$$

The Euclidean distance metric can provide details on pattern matches with respect to magnitude but will fail to capture signals with a similar trend but a different magnitude. As shown in Figure 4.12, users employ the Euclidean distance metric to find counties that are similar to Maricopa county with respect to the volume of transportation over time. Here the user can tell that San Diego county meets the criteria and is highlighted. However, given that the range and magnitude of the time series may vary while the underlying pattern remains the same, this framework also incorporates two other metrics that are less sensitive to magnitude differences within the data. The first is the Sequential Normalized Euclidean Distance which locally normalizes each individual time series for comparison:

$$dist_{SNE}\left(\mathbf{p},\mathbf{q}\right)=\sqrt{\sum_{i=1}^{n}\left(\frac{q_i-\mu_q}{\sigma_q}-\frac{p_i-\mu_p}{\sigma_p}\right)^2}. \tag{4.2}$$

Here μ and σ represent the mean and the standard derivation of the time series respectively. The second is the Mahalanobis Distance which globally normalizes the time series according to the distribution of all time series in the data set:

$$dist_{Mah}\left(\mathbf{p},\mathbf{q}\right)=\sqrt{\left(\mathbf{p}-\mathbf{q}\right)^T\Sigma^{-1}\left(\mathbf{p}-\mathbf{q}\right)}. \tag{4.3}$$

Here Σ is the covariance matrix of the the time series group.

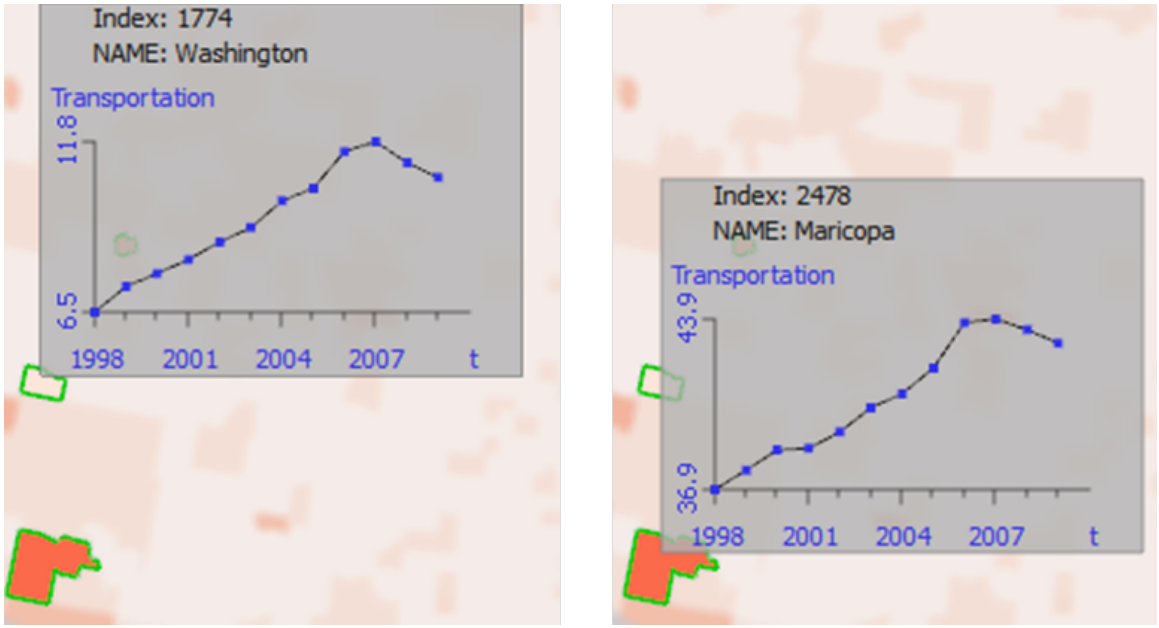


Figure 4.13: An example of time series similarity computed with Sequential Normalized Euclidean distance. Here one can tell the range of temporal values are quite different, but the trajectory shape of the temporal trends is similar.

Sequential Normalized Euclidean distance is useful for finding similar temporal trajectory patterns, regardless of the signal magnitude. As shown in Figure 4.13, while the magnitude of the time series of transportation between Washington county and Maricopa county is quite different, they have very similar temporal trends, thus the similarity criteria using Sequential Normalized Euclidean distance highlighted Washington county. Mahalanobis distance takes the distribution into account making it more suitable for non-uniformly distributed data.

As previously mentioned, in order to calculate a temporal similarity, a reference time series p is needed to provide the basis for comparison. Hochheiser and Shneiderman's Timebox work [105] presented a Timebox widget that can be used to directly specify query constraints on time series data. Inspired by this work, and for the purpose of customizing the similarity parameters in an understandable manner, a

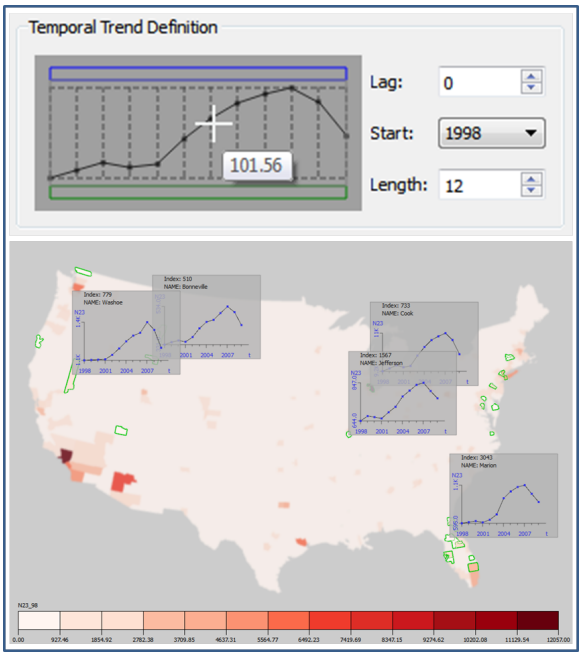


Figure 4.14: The temporal trend definition widget allows users to interactively customize temporal attributes including the temporal value, time series length, lead and lag. Top part of the figure shows the configuration of the temporal trend definition widget and bottom part of the figure shows the filtered result using temporal similarity search based on the defined temporal trends.

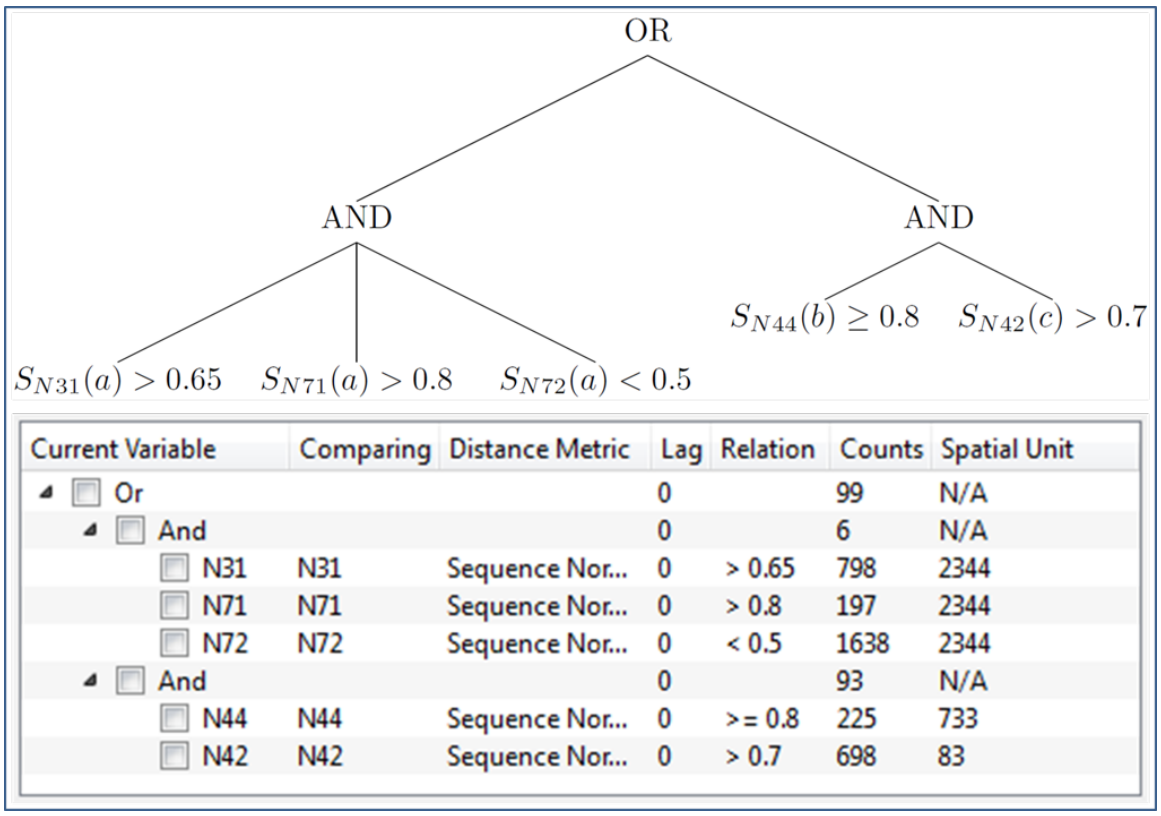
temporal trend definition widget within the similarity widget has been developed to enable users with a flexible search for any trend of interest. As shown in Figure 4.14, the top part of the figure shows a temporal trend and temporal setting for Cook county and a tooltip was shown when mouse move near a time line point. The bottom map highlights the similar spatial units based on the setting in the temporal trend definition widget. The temporal attributes of the time series are customizable, and users can freely edit the time series value by moving the points vertically or changing the minimum and maximum value in the context menu. Indicator lines and value tooltips are provided for auxiliary information on the temporal trend data. The

system also allows users to search for trends by selecting or mousing over a spatial unit under the filtering (brushing) mode, the temporal trend corresponding to the brushed unit will be loaded into the temporal trend definition widget, allowing the user to, for example, look for regions that have similar trajectories to the New York metro area. In this mode, when a user hovers over an area, all other areas that match the similarity criteria will then be highlighted. The top blue bar in this widget indicates the time window for the current temporal variable while the bottom green bar indicates the time window to be compared to. These two bars always begin at the same length. However, when they are dragged or resized, the resulting position will define a temporal shift, thus enabling the analyst to search for lagging or leading temporal trends and identify asynchronous patterns over time.

4.3.2 Logic Tree

While the similarity metrics described above enable exploration between single variables, many space-time processes are a complex combination of multiple variable that require the definition of multiple trends. For example, a user may wish to see where variable A has an increasing trend while variable B has a decreasing trend. In order to support such an analysis given the large number of variables and spatial locations, this framework incorporates the use of a logic widget.

In this manner, the user can refine their queries based on a variety of operations (greater than, less than, AND, OR, XOR). Figure 4.15 shows the logic tree that a user has constructed to search for regions that have a complex set of similarities. The logic tree allows users to dynamically specify a series of filter rules (for example, show me all other regions that have similar temporal trends in their utilities index with respect to my selected region at a similarity value greater than .5 AND show me all other regions that have dissimilar temporal trends in the manufacturing index at a



Current Variable	Comparing	Distance Metric	Lag	Relation	Counts	Spatial Unit
Or			0		99	N/A
And			0		6	N/A
N31	N31	Sequence Nor...	0	> 0.65	798	2344
N71	N71	Sequence Nor...	0	> 0.8	197	2344
N72	N72	Sequence Nor...	0	< 0.5	1638	2344
And			0		93	N/A
N44	N44	Sequence Nor...	0	>= 0.8	225	733
N42	N42	Sequence Nor...	0	> 0.7	698	83

Figure 4.15: The similarity logic tree widget. For simplicity, all similarity measurements used here are the sequential normalized Euclidean metric. Los Angeles county, Cook county and King county are defined as regions a, b,c respectively

similarity value less than .2), where each rule consists of a similarity measurement and a relationship. There is no theoretical limitation on the number of rules that can be combined. However, the result would quickly escalate to null when stacking AND operators.

As an example, let $S_t(i)$ denote the similarity measurements from region i with respect to temporal variable t. In the context of understanding and comparing the industry profiles of different regions, a user could explore which areas' manufacturing similarity scores are larger than 0.65 and the arts & entertainment similarity scores is larger than 0.8 but the accommodation similarity is less than 0.5 when compared to

region a. The analyst may also want to determine which areas' wholesale similarity is larger than 0.7 compared to region b but the retail similarity is not less than 0.8 when compared to region c. Such a query would result in the following tree and its corresponding layout in the logic widget as shown in Figure 4.15. Each branch of the logic tree can be interactively visualized through choropleth map highlighting so that users can quickly explore a variety of combinations. Each tree is directly associated with a single areal unit in the map, and is stored in memory so that if users were to select a different region of interest and then return, their previously created logic tree would still be associated with that region until the user chooses to delete it. In this manner, this framework provides an alternative to the traditional means of searching for similar spatial trends by animating a choropleth map or utilizing linked views. In this scheme, users may now query the data through interactive widgets to directly search for comparable trends. From there, they can animate the map which has now applied the previously mentioned focus+context blurring to the choropleth maps so that attention can directly be focused on the filtered areas which are known to have similar temporal trends.

4.3.3 Temporal Trend Multiples

To provide overview + detail views for the temporal similarity brushing results, a novel temporal trend multiples vies for displaying and ordering time series of all brushed spatial units has been developed. The temporal trend multiples are ordered in a row-wise manner where the most similar time series is positioned in the left top corner (Figure 4.16). This matrix layout is similar to the work by Turkay et al. [145]. Moreover, this view is enhanced such that the background of the multiples is shaded by the geographical distance to the user selected geographical region. By matching

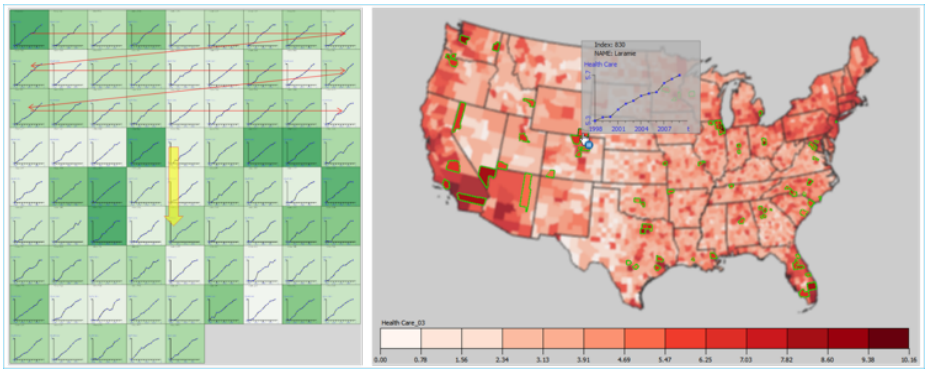


Figure 4.16: Temporal trend multiples. (Left) The red lines show the exact arrangement of the multiples based on the similarity to the temporal trend of user's interest. In this case the ordering is row-wise meaning that similarity decreases row by row. The ordering is denoted by the red arrows. The background color of a plot encodes the spatial proximity. In this example, the user explores Health Care trends in Laramie County using a similarity threshold of 92%.

whether the colors of multiples have a certain trend, users can assess the association between spatial closeness and temporal similarity.

4.3.4 Comparison to Previous Work

It is critical to note that the concept of linking analytical methods to brushing has been explored for multivariate data. Bernard et al. [176] addressed the problem of interactive search in time series and multivariate data with a visual catalog of time series data and content-based visual query specification. Hao et al. [177] also introduced the Intelligent Visual Analytics Query that allows interactively selecting focus and identifying the relationships to other portions of the data set. However, this framework has several advantages:

1. Previous systems do not provide a geographical context. When projecting the temporal similarity to the geographical space, highlighting and exploration be-

comes problematic as users must compare not only the temporal trend, but also the spatial extents.

2. Previous systems lack the capability of restricting temporal comparisons to specific local areas based on the geographical properties, such as only the first order neighbors.

3. This framework visualizes the details of the similarity criteria via the distance map. This is critical for improving an understanding of how temporally similar regions are distributed in geographical space.

4. Previous systems do not integrate the lag and lead concepts into the similarity search which is important for developing hypotheses on causal drivers. While techniques such as dynamic time warping [110] are able to provide a metric of similarity between time series that are not necessarily aligned, our domain experts indicated the desire to define and explore lags and leads as a means for hypothesis generation.

5. While other systems perform data reduction methods for clustering of time series patterns, methods such as principal component analysis or multi-dimensional scaling are often difficult for users to understand. As such, this framework opts for allowing the user to specifically define their search criteria. The trade-off is a more complex input where, in the case of data reduction, the result may be a more complex output.

4.4 Multivariate Similarity

While identifying similar temporal patterns is a critical analytical task, the goal of this thesis is also to enable users with the capability of exploring multivariate

similarity in geographic space. Recent work in the visual analytics community has focused on multivariate analysis and distance function definitions for similarity analysis [159, 178, 179], yet little work has been presented on exploring these multivariate projections in geographic space or providing users with insight into the actual distance in the multivariate space. The developed framework utilizes two methods for exploring multivariate similarity. One is the multidimensional distance map, the other is multivariate clustering.

4.4.1 Multidimensional Distance Map

Often times, analysts may wish to know how similar regions are with respect to a set of multiple measures (e.g., are these regions similar in terms of crime, health, and economic characteristics?). However, in geographical space, the locations of spatial units are predefined and a rearrangement of those units would lead to difficulties in analysis. The multivariate distance map in this framework provides analysts with a means of defining the multivariate distance in the geographical projection space. The multivariate similarity brushing ties the user defined similarity metric (e.g., Euclidean distance) to the mouse pointer. As the pointer touches a geographical area, the map is recolored to represent the distance from all other geographical areas to the selected unit (Figure 4.17).

In this way, users can maintain a relative idea of the actual distance in the multivariate space such that they can determine the compactness of a region with regards to a multivariate distance function. The benefit of visualizing the multivariate distance function (in conjunction with directly visualizing the projected clusters) is that by directly mapping the distance function to a choropleth map, users can now also ascertain the relative "closeness" of geographic areas within the multivariate space.

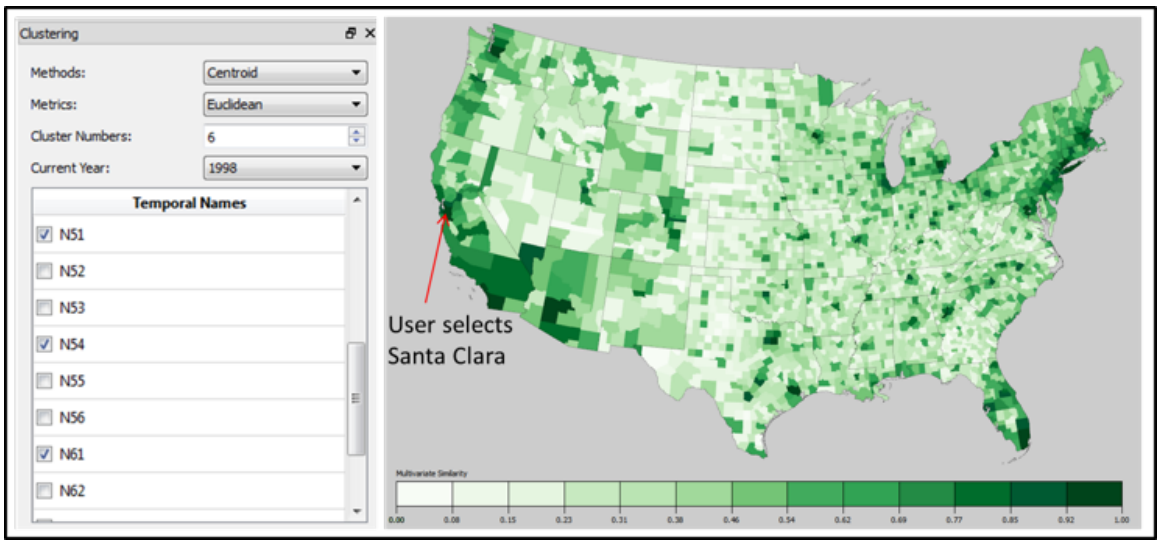
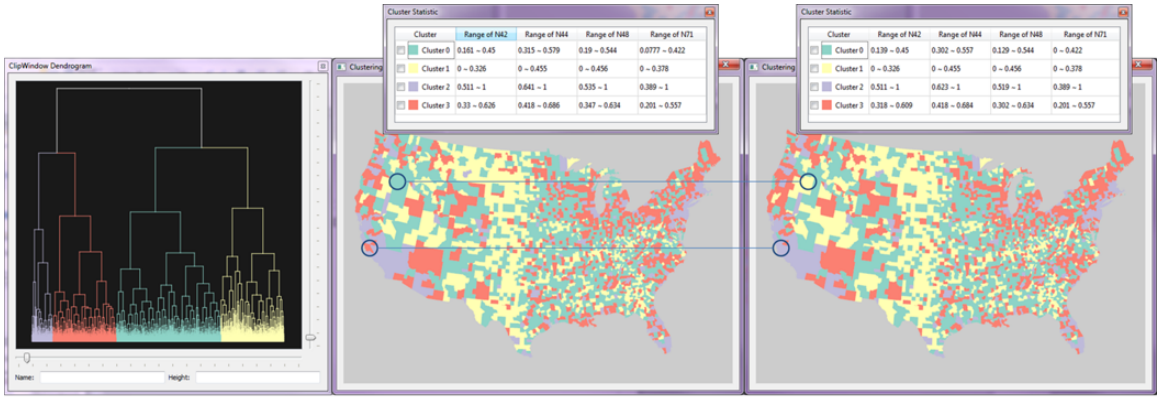


Figure 4.17: Here the user is exploring the effects of applying the multivariate similarity tooltip. In the leftmost screen, the user defines the multivariate distance function selecting Information based services, Educational Services and Professional Service metrics. The user explores which regions of the country are similar to Santa Clara County (Silicon Valley) in 2007 (a known knowledge economy). Regions in dark green are most similar (closest in the multivariate Euclidean measure). Here one can quickly see the knowledge economies of the East and West coast as well as localized pockets across the country which may be of interest for future exploration.

4.4.2 Multivariate Clustering

There exists many clustering methods [180], and the use of different algorithms could lead to very different clustering results (Figure 4.18(b)(c)). This visual analytics framework supports various clustering methods (hierarchical clustering, k-means Clustering) along with several types of distance metrics (e.g., Euclidean, Manhattan). By assigning each cluster a unique color and projecting it onto the map, the framework is able to create a clustering choropleth map, Figure 4.18(b)(c).



(a) Dendrogram (b) Hierarchical Clustering (c) K-means Clustering

Figure 4.18: Here is an example that illustrates a dendrogram and the multivariate clustering results using hierarchical clustering and k-means clustering. The circled units are the differences between the two clustering results. However, it is hard to find all of those differences by manually inspecting the two clustering results. The variables selected for clustering are the amount of Retail, Wholesale, Transportation and Entertainment. Using the statistic view, one can tell that there exists correlation between those variables.

This visual analytics framework can be used to assess and manipulate multidimensional spatial clustering, here the k-means clustering algorithm is used in the following sections as an example. The k-means clustering algorithm has been picked because it has remained the core algorithm for the computation of geodemographic classifications due to its efficiency and simplicity [125, 181]. One well known example is the UK National Statistics Output Area Classification (OAC) which is an open geodemographic classification with a hierarchical structure of 7 supergroups, 21 groups and 52 subgroups [120]. This type of multivariate clustering has served as the basis for various geovisualization techniques. For example, Slingsby et al. [182] developed rectangular hierarchical cartograms for mapping socio-economic data of OAC, and also proposed a set of interactive visualization techniques to explore pop-

ulation profiles of areas and how uncertainty in OAC varies geographically and by OAC category [183]. As for k-means, the result is sensitive to initial centroids, i.e., different initiations would make a big difference in the outcome. However, there is no single best approach for selecting the best clustering algorithm, just as no clustering algorithm offers any theoretical proof of its certainty [184]. Thus, achieving a reasonable clustering result has always been a dilemma. To overcome such issues, this framework provides a ranked list based on Davies-Bouldin index (DBI) [185] to allow users to pick a good base clustering result for future exploration. The DBI is an internal evaluation scheme that can assess the quality of clustering. It is defined as:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}(\frac{S_i + S_j}{M_{i,j}}) \tag{4.4}$$

where n is the number of clusters, S_i is the measure of scatter within the ith cluster, and $M_{i,j}$ is the measure of separation between the ith cluster and jth cluster. From the definition, the smaller the DBI, the better the clustering solution. In this framework, every k-means clustering will run multiple times for the same cluster number k with different initial centroids. Some of the runs will initialize centroids by sampling from the original data points, others will initialize the centroids from the result of other clustering algorithms (e.g., hierarchical clustering) as it can help achieve better clustering results. At the end of the process, the framework will sort and display the top ten models with the lowest DBI generated from all the runs (Figure 4.19).

4.5 Interactive Clustering

Due to the often non-intuitive connection between multivariate space and geospace, it is a challenge to simultaneously explore the clustering result in both multivariate space and geographical space. As noted in the related work, clustering with feedback is one of the most important aspects for multivariate data analysis in geospace.

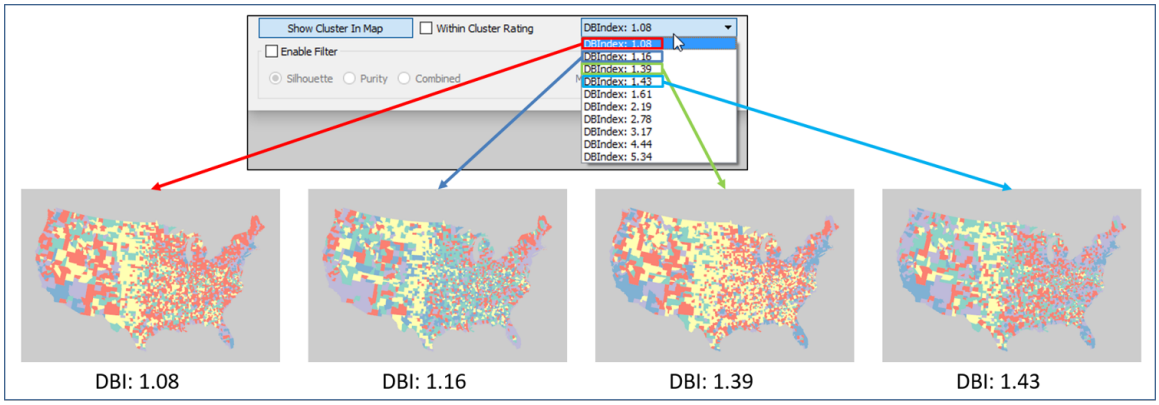


Figure 4.19: An application of the DBI value in the framework after running the same clustering algorithm multiple times. When users select a certain DBI value from the drop-down list, corresponding clustering result is displayed in the map widget.

Thus, the framework provides interactive clustering that can adapt to the spatial constraints (e.g., users can identify potential spatial units that may need to be re-categorized into other clusters and manipulate the clustering results). Dynamic filters with three criteria are also provided to help users assess the clustering results. The first criterion is the silhouette coefficient [186]. The silhouette coefficient refers to a method of interpreting clusters which allows users to know how well each object lies within its cluster. This coefficient is used as a criteria for discovering the border point in multidimensional space and defined as:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, \tag{4.5}$$

where $a(i)$ is the average dissimilarity of object i with all other objects within the same cluster, and $b(i)$ is the lowest average dissimilarity of object i to any other clusters where object i is not a member. From the above formula it is clear that $-1 \leq S(i) \leq 1$. When $S(i)$ is close to 1, it means the datum is appropriately clustered. When $S(i)$ is close to -1, it means object i would be more appropriate labeled if it was clustered in one of its neighboring clusters. When $S(i)$ is close to zero, it means that the

datum is on the border of two natural clusters. Therefore, this framework leverages this coefficient as a means of assessing the boundary elements between clusters. The objects with values less than or equal to zero are usually our points of interest.

Users may also be interested in the relativeness of cluster labels within a neighboring area. The Gini index-like [187] purity indicator is the second criterion for geographic multivariate clustering inspection. The purity indicator is defined as:

$$P(i) = (\frac{n_{C_i}}{N(i)})^2 - \sum_{C_i \neq C_j} (\frac{n_{C_j}}{N(i)})^2, \tag{4.6}$$

where n_{C_i} is the number of units that belong to the same cluster of i (C_i) in i's neighborhood, n_{C_j} is the number of units that are different from i's cluster in i's neighborhood, $N(i)$ is the total count of units of i's neighbors, and $P(i)$ also lies in the range: $-1 \leq P(i) \leq 1$. When $P(i)$ is close to 1, it means that unit i is almost surrounded by the neighbors within the same cluster. When $P(i)$ is close to -1, it means that unit i is almost surrounded by the neighbors from another cluster. When $P(i)$ is near zero, it means its neighbors are randomly scattered in different clusters. Thus, the higher the purity value a unit has, the stronger the spatial association is around that unit.

Based on the silhouette and the purity criterion, the local point of interest is the third criterion developed as a combination of both. The local point of interest is defined as:

$$LPOI(i) = \frac{n_{C_i}}{N(i)} * \frac{a(i)}{a(i) + L_{C_i}} - \sum_{C_i \neq C_j} \frac{n_{C_j}}{N(i)} * \frac{a(i)}{a(i) + L_{C_j}} \tag{4.7}$$

where L_{C_i} is the average dissimilarity of object i with all other objects within its same cluster in i's neighborhood, and L_{C_j} is the average dissimilarity of object i with objects that are in other clusters in i's neighborhood. The range of $LPOI(i)$ is -1 to 1 as well. Values near 1 indicate the unit is stable as it is surrounded by similar

units all in its same cluster. When the value is closer to -1, the unit is surrounded by similar units but all from another cluster, indicating this unit may be a candidate for relabeling.

All three indices can be displayed in the thumbnail plot when users hover the mouse over a certain spatial unit. By setting the upper and lower bound of a certain criteria, the corresponding units filtered out will be highlighted in dark color contours for the user to inspect.

The framework implements two modifications for flexible direct manipulation. One method is result manipulation, which means the modification will only affect on the label index of the user selected units. The other method is model manipulation which will affect the weights in the clustering process and eventually the labels of other data points. Each element in the dataset will have an associated weight which can be modified through user interaction. Suppose there are n units $u_1, u_2, \ldots, u_n$ and their weights are formed as $w_1, w_2, \ldots, w_n$, such that initially each spatial unit u_i will have the same instance weight $w_i = 1$ influencing the placement of the centroids. After the initial clustering, analysts may assign unit u_i to specific cluster C_j, then u_i's weight will be modified to be either based on the cluster size s_j such that $w_i = 1+\sqrt{s_j}$ or a predefined constant value such as $w_i = 1+const$. Thus, during each iteration, u_i's proximity to C_j's centroid c_j will be computed by multiplying that weight w_i, so it is more likely to be assigned to cluster C_j. When calculating the cluster centroid, u_i will only contribute its weight to that cluster. The new centroid of C_j will be $c_j = \frac{\Sigma_{k \in J} w_k u_k}{\Sigma_{k \in J} w_k}$ where J is the set of unit indices which have been assigned to C_j. Eventually, this result in cluster C_j's centroid c_j moves towards u_i and this unit will likely belong to that cluster (Figure 4.20). Based on the definition of direct manipulation [188, 189], the modified result will be synchronized with all the visual widgets simultaneously.

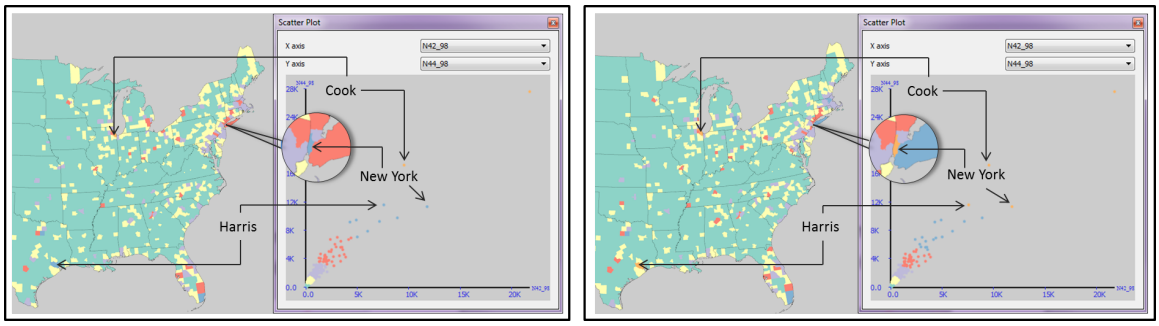


Figure 4.20: An example of user-guided k-means: The left figure shows a projection of the k-means clustering (where k=6) of counties based on their measured indices of wholesale trade, retail trade, entertainment and accommodation. These clusters are projected onto a choropleth map and the analyst sees that Cook County belongs to the orange cluster and New York County belongs to the blue cluster. Based on their domain knowledge of the data, the analyst knows that Cook county and New York County should actually belong to the same cluster, thus the analyst interactively reassigns New York County to the orange cluster. The resultant weights matrix is updated and k-means clustering is recomputed using the user selected weights. The resultant clustering not only modifies Cook and New York Counties, but it also causes other counties to be reassigned to different clusters, such as Harris County being reassigned to the orange cluster. The scatterplot view is also shown (projecting into the retail trade vs. wholesale trade space) along with each map to illustrate the 2D clustering of the data in the multidimensional space.

Chapter 5

CLUSTERING EXPLORATION

This thesis links users to a variety of clustering quality measurements to enable them to develop an understanding of global and local multivariate clustering for geographical visualization. It supports the visual exploration of geographical projections of both k-means and hierarchical clustering. To address important differences in local trends related to either spatial dependence or spatial heterogeneity, this thesis characterizes space into the following four categories:

- *Discrete spatial extent* – Particular types of data may be reported in such a way that they are bounded by a fixed spatial extent. Previous research tends to apply multivariate clustering to the entire spatial extent, which can conceal local variations. This framework enables geographical selections at a constant spatial extent and allows users to apply multivariate clustering to the selected spatial extent. The interaction and visual encoding enables users to identify local patterns between places in order to understand the impact of the spatial heterogeneity on the multivariate clustering procedure.

- *Discrete geographical features* – Different geographical features are not always spatially continuous. Thus, this framework allows users to distinguish geographical features (e.g., urban vs. rural) and then apply multivariate clustering to the geographical features of interest.

- *User-defined (continuous) spatial extent* – While many geographic studies examine phenomenon where the spatial extent is fixed, many other questions require

an analyst to modify the spatial extent by zooming in to a particular set of spatial units or zooming out to a particular extent, and then performing a cluster analysis. In this context, the arrangement and the neighborhood structure of the data are variable. This framework allows users to adjust spatial scales around a fixed location to understand the impact of the spatial dependence on the multivariate clustering procedure.

- *Continuous geographic resolution* – Another issue to consider when evaluating multivariate clustering results in geographic space is the impact on the results of varying the resolution of the data. This variation in the resolution of the spatial units of interest is otherwise known as the problem of modifiable areal units [190]. It is known that using larger areal units (i.e. states as opposed to counties) reduces the variance in the data [191]. This framework allows users to aggregate multivariate attributes at different spatial resolutions (e.g., county, state) to understand the impact of the spatial dependence on the multivariate clustering procedure.

5.1 Group Selection

To enable the exploration of the spatial impact on multivariate clustering, this framework extends the traditional selection operation through the concept of group operations. Three types of selection are fully implemented in the framework:

- Rubber band selection in geographic space;
- Selections from multivariate space utilizing histogram, scatterplot, categorical view, etc;
- Automated geographical selections such as selecting based on a boundary layer or using a neighborhood.

These three selection methods enable users to define any desired areas. The group level operations include updating all the exploratory data analysis widgets (e.g, histogram, scatterplot), applying local clustering, and aggregating local clustering statistics.

5.2 Visual Exploration Widgets

In addition to the PCA scatterplot introduced in Section 4.2.5, this framework provides several other visual analytics methods for clustering exploration including the dendrogram, statistic table, PCP (Parallel Coordinate Plots) area profiler, and rose plot.

Dendrogram: For each hierarchical clustering, an interactive dendrogram (Figure 4.18(a)) is generated. This allows users to freely navigate and select any branch for highlighting units on the map.

Statistic Table: All the clustering results will derive a statistic table view that illustrates the range of each variable for each cluster. From the statistic view analysts can identify interesting properties of the clusters (Figure 4.18).

PCP Area Profiler: While PCA is good for visualizing the multivariate distance, it lacks consistency in appearance as the data change. Oftentimes, it is not enough to determine the cluster label just by using the PCA scatter plot when the border point is at the intersection area of more than two clusters. Hence a PCP area profiler has been implemented to visualize the multivariate relations of different area profiles in a simple click. There is one customizable area profile where users can select the units of interest and three predefined area profiles: the local neighboring area which only considers the units within the first order contiguity of the selected unit; the intra-cluster area which only considers the

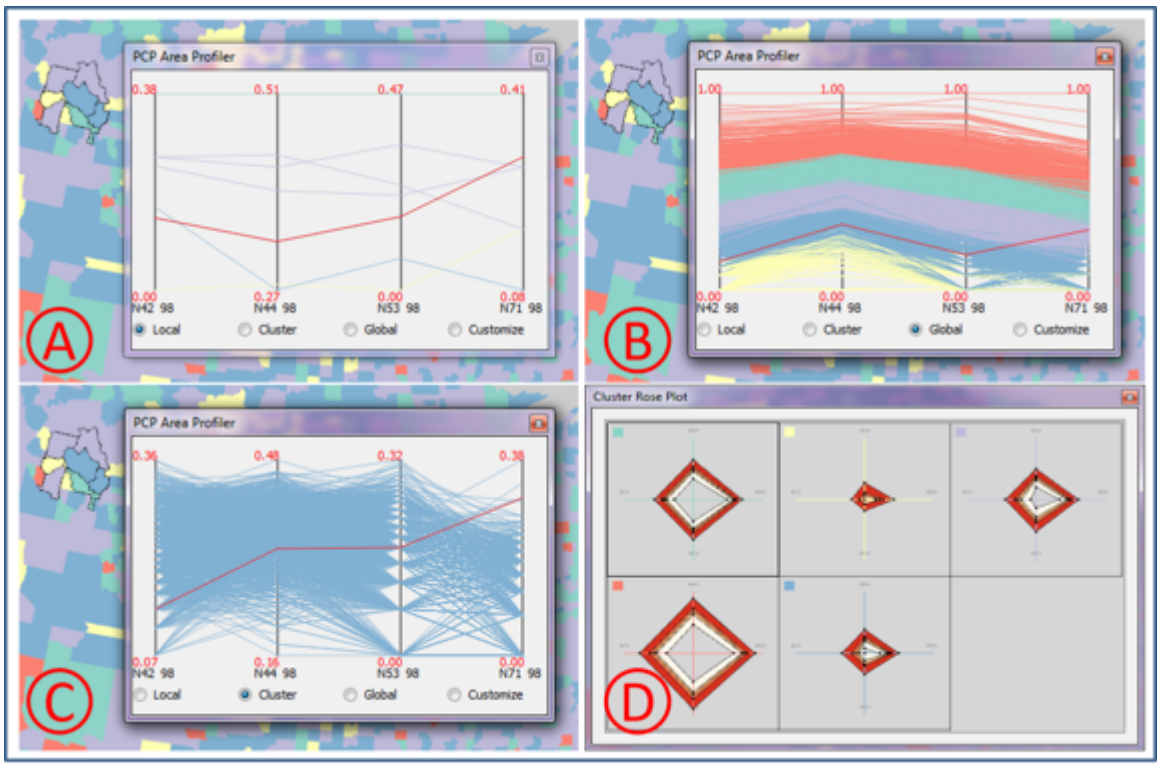


Figure 5.1: An example of the PCP area profiler and Rose plot. Here users explored the PCP area profiler with three different profiles. The Rose plot in the bottom-right shows the clustering result of five clusters, which allows users to learn about the characteristic of each cluster.

units in the same cluster as the selected unit, and; the global area which considers all the units. By switching among those area profiles, users can explore how the datum is distributed in the multidimensional space (Figure 5.1(A-C)). When a user mouses over the unit, the corresponding unit lines in the PCP will be rendered in red for easy identification.

Rose Plot: While PCPs provide a detailed view of the data values, they are often very cluttered. This framework employs a modified version of the traditional rose plot (Figure 5.1(D)) akin to Schreck et al. [192]. Each variable axis has five

points which indicate the lower bound, three quartiles, and the upper bound of each variable respectively. While Gestalt principles note that humans are good at shape comparison, drawbacks of the rose plot include shape changes due to axis ordering and the often unintuitive scaling that must be done per axis.

5.3 Clustering Comparison

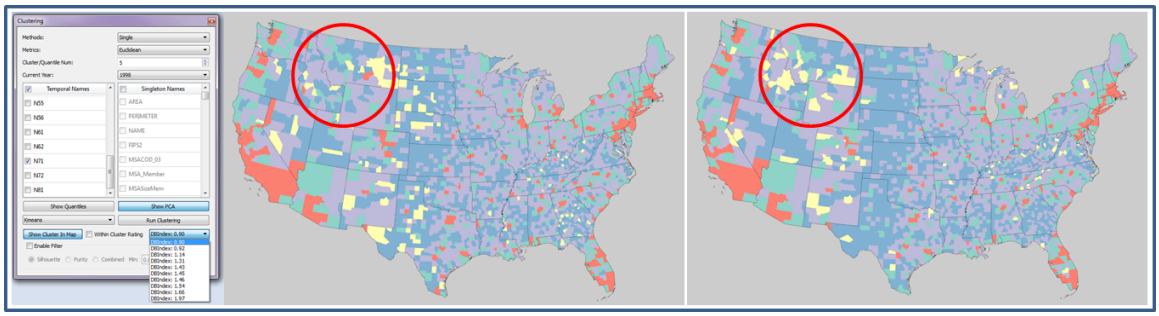


Figure 5.2: The coherent clustering color mapping with both maps having five clusters. By maintaining label consistency for generalized clustering comparison, users can quickly tell that the clustering results are similar while at the same time noting that there exists differences in the northern part of the US (the red circle). However, it is still difficult for users to figure out exactly how many differences there are.

Even though many clustering comparison methods have been developed, designing an effective comparison operation and showing the results is still a challenge. Hu et al. [136] proposed the coherent clustering color mapping that attempts to keep cluster labels of spatial units consistent between different clustering results. They assign the same label (color) to the clusters with the maximum number of correspondences to facilitate the comparison of clusters. To enhance the coherence and generalized cluster comparison between the multiple runs k-means clustering results, the framework implements a coherent clustering color mapping algorithm similar to the work from Hu et al. [136]. The coherent clustering color mapping attempts to keep cluster

labels of spatial units consistent between different clustering results. For example, if for clustering C^1 there exists two clusters C_1^1 with spatial units $\{a, b, c\}$ and C_2^1 with spatial unit $\{d\}$. While for clustering C^2 there exists two clusters C_1^2 with spatial units $\{c, d\}$ and C_2^2 with units $\{a, b\}$. Then the process is try to minimize the number of spatial units that will be relabeled between them. Thus, in this example, one can say that C_2^1 is changing to C_1^2 and C_1^1 is changing to C_2^2. In order to determine this minimal label change, the Kuhn-Munkres (KM) algorithm [193] is implemented. The KM algorithm treats the clusters between different time steps as a weighted bipartite graph and solves it as an optimal assignment problem. Assume the clusters in one clustering are set X such that $X = \{x_1, x_2, \ldots, x_n\}$, and the clusters in another clustering are set Y such that $Y = \{y_1, y_2, \ldots, y_n\}$, where both time steps have an equal number of clusters. Then the weight is defined using a cost function (closeness measurement):

$$c(x_i, y_j) = |x_i| + |y_j| - 2 * |x_i \cap y_j| \tag{5.1}$$

and the objective function is equivalent to finding a permutation π of $1, 2, \ldots, n$ such that $\Sigma_{i=1}^{n} c(x_i, y_{\pi(i)})$ is minimum.

By executing the KM algorithm to solve this objective function for every two adjacent clustering results in the rank list, the same label (color) can be assigned to the clusters with the maximum number of correspondences and eventually facilitate the intuitive comparison of the spatial clusters. However, this method has several drawbacks: first, it lacks a detail comparison capability (in Figure 5.2, users can not tell the exact difference between two clustering results when the amount of spatial units or the number of clusters is large); second, it can only compare clustering with the exact same cluster numbers and units, and; third, it is not genuinely coherent because the process does not have the transitive property. To overcome these limitations, a novel visual analytics tool called the Triple-D View (Drag and Drop clustering Dif-

ference View) was developed to simplify the process of clustering comparison in this framework. By dragging and dropping between two clustering results, the view not only can visualize the difference between clustering results regardless of the cluster number and coloring scheme but can also generate a numerical index to help users assess the clustering similarity.

Each cluster is essentially a set, thus comparing the difference between clusterings is equivalent to exploring the changes among those sets. According to observations, the changes have been generalized into a combination of the splitting step and the merging step. To keep the idea simple, consider the example in Figure 5.3. Here, we have a clustering result Ω of 3 clusters A, B, C for 15 objects on the left, and another clustering result Ω' of 4 clusters A', B', C', D' for those same 15 objects on the right. In the splitting step, we subdivide Ω into small clusters. For instance, for cluster A, objects $1, 3$ are formed into the same cluster in Ω', objects $4, 6, 7, 9$ are formed into the same cluster in Ω', and object 14 is merged into another cluster in Ω'. Thus we will have three sub-clusters in Ω'' for cluster A. In the merging step, we just need to check each cluster in Ω' to find out which small sub-clusters in Ω'' it contains. The intermediate clusters are actually the mutual information between these two clustering results. This process is demonstrated in Figure 5.3.

By dragging one clustering result and dropping it onto another clustering result in the Triple-D view, the Triple-D view will map the changes (i.e., the intermediate sub-clusters) under the two clusterings being compared (Figure 5.4). The layout of our difference view is an inverted pyramid which is similar to the GTdiff method [108]; however, GTdiff only provides comparison for temporal bins as a difference of values between time steps. Here, this layout is utilized to represent the difference between different clustering results. To represent the changes, three criteria for the proportion are defined: less than 50 percent, larger than 50 percent, and equals to 100 percent.

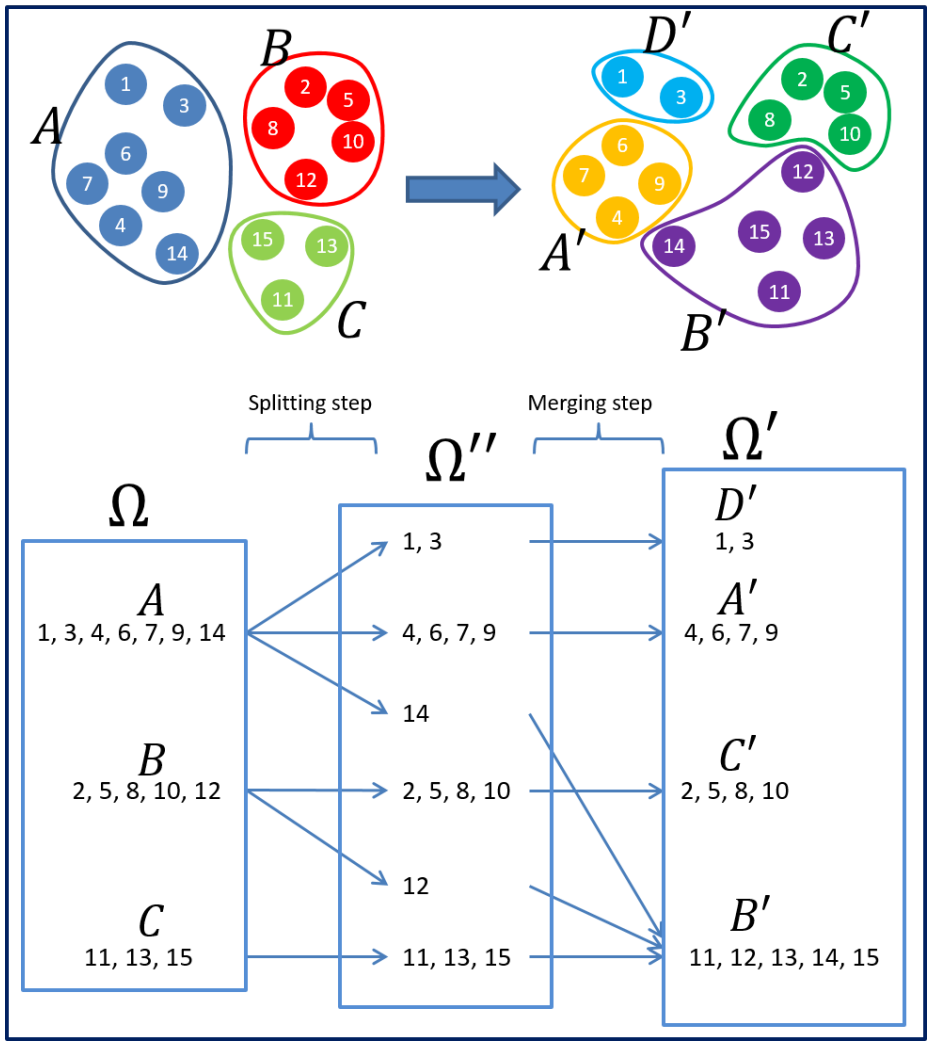


Figure 5.3: Comparing two clustering results for the same group of 15 objects. In the top figure, the left part is a clustering result Ω with three clusters, and the right part is a clustering result Ω' with four clusters. The bottom figure is the illustration of the comparison process. The value on the arrow indicates the proportion of the sub-cluster in that step.

The proportion in the splitting step refers to the ratio between the size of the sub-cluster and the size of its original cluster where the sub-cluster splits from, in the merging step it refers to the ratio between the size of the sub-cluster and the size of

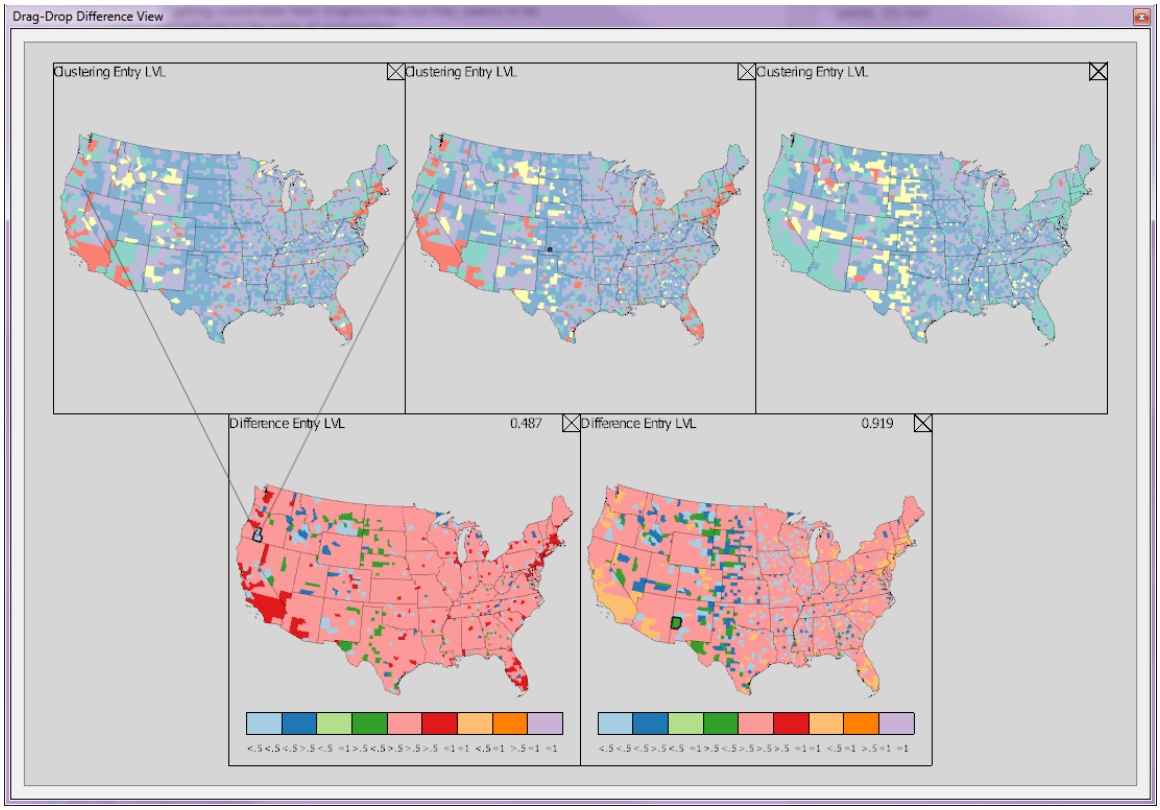


Figure 5.4: An example of the Triple-D View (Drag and Drop clustering Difference View). The top three maps are three different clustering results using k-means but with different initial centroids respectively. The bottom two maps are the comparison results of the first two and last two respectively. When users click on a certain unit in the comparison result, indicator lines will be drawn on top of them to mark the corresponding units from the two compared clustering results.

its successive cluster which the sub-cluster merged into. As there are three criteria for both steps, there will be 9 variables that can be used to represent the changes. An example of the visual coding is shown in Figure 5.4.

The Triple-D view not only visualizes the difference between clustering results regardless of the cluster number and coloring scheme, but it also generates a numerical proximity metric that obeys all the metric properties (positivity, symmetry, triangle

inequality, indiscernibility) to help users assess the clustering similarity. In contrast, the Rand Index is not suitable for unlabeled clustering comparison as it requires a ground truth, and NMI (Normalized Mutual Information)/Variation of Information can not handle the situation when mutual information is 0, and the similarity measure introduced by Torres et al. [194] does not provide diversity/entropy information for the comparison which make the result less meaningful. So the split-merge metric in the framework is defined as:

$$SM(\Omega, \Omega'') = -\sum_{i}\sum_{j} \frac{|C_i \cap C_j'|}{N} log \frac{|C_i \cap C_j'|^2}{|C_i||C_j'|}, \tag{5.2}$$

where Ω and Ω'' are the two clustering results been compared. C_i is the ith cluster in Ω, C_j' is the jth cluster in Ω'', and N is the total number of units. Larger metric values represent more dissimilarity between clusterings.

Chapter 6

VISUAL APPEARANCE OF SPATIAL ASSOCIATION

The analysis and understanding of spatial patterns is essential to all subfields of geography [195], and the visual representation of spatial patterns represented is greatly affected by the choice of classification boundaries. The current methods introduced in this framework have focused on the identification and exploration of class members who fall near a classification boundary. In this section, the focus is on quantifying the visual impact that changing the label of an element will have. As such, there are three types of spatial association patterns that can be identified:

- *Clustered:* Map elements with the same label are contiguous in geographic space, as indicated by positive measures of spatial autocorrelation in Moran's I.
- *Dispersed:* Map elements with different labels (but with a repeated pattern) are contiguous in geographic space, as indicated by negative spatial autocorrelation in Moran's I, an example of such a pattern would be a checkerboard.
- *Random:* Map element labels are randomly distributed on the map, as is indicated by a Moran's I near zero, i.e., the distribution of regions with similar properties is unspecified/random in geographic space.

Each type of pattern is associated with a description of the visual appearance of the map, and these spatial association patterns are typically defined and tested using spatial autocorrelation. Spatial autocorrelation is often used with p-value, z-score, and resampling methods (e.g., Monte Carlo sampling [196], randomization tests [29]) to indicate the significance level of the tendency of spatial clustering in a map. Given

that visual appearances are directly related to measures of spatial autocorrelation, our goal is to adapt an indicator of spatial association to quantify the visual change that may occur in a choropleth map as an element's class label is altered.

Many indicators for spatial association exist (e.g., join count statistics [197], Geary's C [81], Moran's I [198], Getis-Ord General G [82]). However, these statistics are all special cases of cross-product statistics [199, 200]. Moran's I [198] is perhaps the most well-known and widely used measure of spatial autocorrelation. Moran's I is defined as:

$$I = \frac{N}{\Sigma_i \Sigma_j w_{ij}} \frac{\Sigma_i \Sigma_j w_{ij}(x_i - \bar{X})(x_j - \bar{X})}{\Sigma_i (x_i - \bar{X})^2}, \tag{6.1}$$

where N is the number of spatial units indexed by i and j, x is the variable of interest, $\bar{X}$ is the mean of x, and w_{ij} is an element of a matrix of spatial weights.

Unfortunately, Moran's I is designed for continuous variables. Since the visual appearance of the map relates solely to the final class labels, we need a metric that can be applied to categorical data values. As such, we modify the Moran's I measure to provide a metric of spatial autocorrelation based on the class labels. To do this, we need to redefine the variables in Equation (6.1). x_i is now defined as a vector $(c_1, c_2, \ldots, c_n)$, where n is the number of clusters and c_n is a binary value, 0 or 1, such that if element i belongs to cluster 1, then $c_1 = 1$ otherwise, $c_1 = 0$. Then $\bar{X}$ will be the average of all vectors x_i, and a modified global Moran's I can be calculated to evaluate the spatial association of cluster labels. Note that in this paper, we utilize the Queen contiguity for defining the spatial weights matrix. And $w_{i,j} = 1$ for all Queen contiguous neighbors in our implementation. While the choice of the spatial weights matrix will impact the calculation, the application is generalizable to any spatial weights choice.

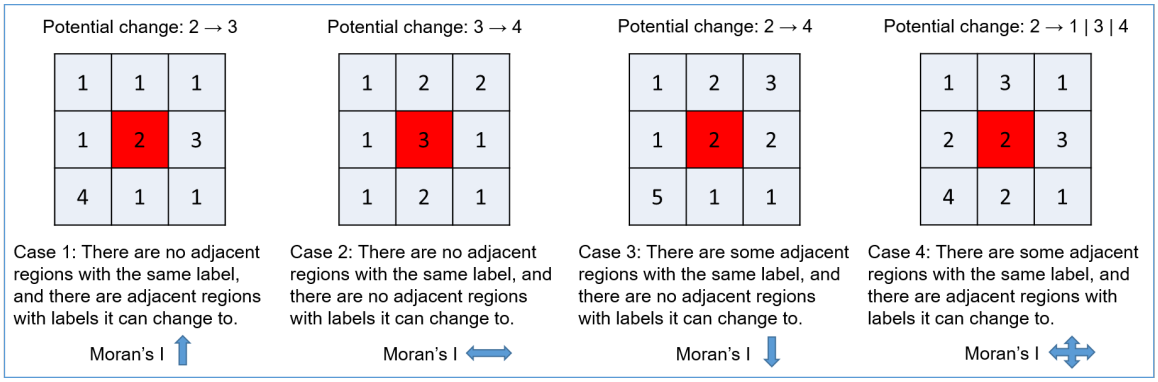


Figure 6.1: Four spatial cases and the effects of changing a single unit.

6.1 Categorizing the Effects of Relabeling

Once a measure for the spatial association of the class labels is defined, the next step is to determine the cases in which altering a label will impact the visual spatial association. We identify four potential spatial arrangements for elements on a choropleth map, Figure 6.1. Based on these arrangements, we then define the value change in our modified Moran's I that would result in a change of the classification label.

Case 1 The spatial unit under analysis, i, is spatially contiguous only to units with different class labels. The position of i in the classification space is such that it lies near the class boundary of one or more spatially contiguous units. In this case, if i was relabeled, the spatial association will increase. This is illustrated in Figure 6.1 (Case 1). Here, unit i is the red square and has a class label of 2. This element lies near the boundary of class 2 and class 3. If the label of i were to change from 2 to 3, an increase in visual clustering could be observed and the spatial association value would increase.

Case 2 The spatial unit under analysis, i, is spatially contiguous only to units with different class labels. The position of i in the classification space is such that it does not lie near the class boundary of any spatially contiguous units. In this

case, if i was relabeled, there would be no change in the spatial association. This is illustrated in Figure 6.1 (Case 2). Here, unit i is the red square and has a class label of 3. None of its neighbors share the same class label, thus i does not add to any visual cluster. i lies on the boundary of class 3 and class 4; however, changing i's label to 4 does not result in i visually combining with other spatially contiguous regions, thus there is no change in the spatial association metric.

Case 3 The spatial unit under analysis, i, is spatially contiguous to some (or all) units which share the same class label. The position of i in the classification space is such that it does not lie near the class boundary of any other spatially contiguous units. In this case, if i was relabeled, the spatial association will decrease. This is illustrated in Figure 6.1 (Case 3). Here, unit i is the red square and has a class label of 2. Several of its neighbors share the same class label, thus forming a small region that will visually appear to be clustered. While i does lie near the boundary of class 2 and class 4, there are no spatially contiguous elements labeled class 4. As such, if i were to be relabeled, the size of the region containing elements with a class label of 2 would decrease, and no other region in this scenario would add i to their spatial grouping. As such, the visual clustering would decrease, resulting in a lower spatial association value.

Case 4 The spatial unit under analysis, i, has a label which lies near a classification boundary and is spatially contiguous to some units which share the same class label. The position of i in the classification space is such that it does lie near the class boundary of other spatially contiguous units. In this case, if i were to be relabeled, the change in spatial association could be positive, negative, or neutral depending on the number of contiguous units (and their contiguous

units) that have the same label as i. This is illustrated in Figure 6.1 (Case 4). Here, unit i is the red square and has a class label of 2. Several of its neighbors share the same class label, thus forming a small region that will visually appear to be clustered. However, i lies on the boundary of class 2 and class 4 and is spatially contiguous to other regions with a class label of 4. If i were to be relabeled, the size of the region containing elements with a class label of 2 would decrease; however, the size of the region containing elements with a class label of 4 would increase. As such, the modified Moran's I would need to be recalculated for the entire map to determine the net change in spatial association.

While Cases 1-3 are straightforward to identify, Case 4 is perhaps the more common case in choropleth map design. Thus, for a unit i in Case 4, we define the number of regions that belong to the same cluster as i in its surrounding area as p_i. The number of regions that belong to a cluster that i can change to in its surrounding area as q_i. The effect on the spatial association after i is changed is based on the number of surrounding units that i can change to and is proportional to $q_i - p_i$. Figure 6.1 only considers the effect of a single changeable unit, we extend this to more complex situations (Figure 6.2(A)) in which several contiguous regions could change, resulting in a cascade of visual clustering patterns.

Theorem 1 *If the potential changeable regions are not adjacent, then their effects on the spatial association are separate/independent.*

By inspection, one can observe that if spatial units that are identified as having class labels near boundaries are non-adjacent, then the effect of modifying their class labels will be independent. This can be observed in Equation (6.1) where units that are not

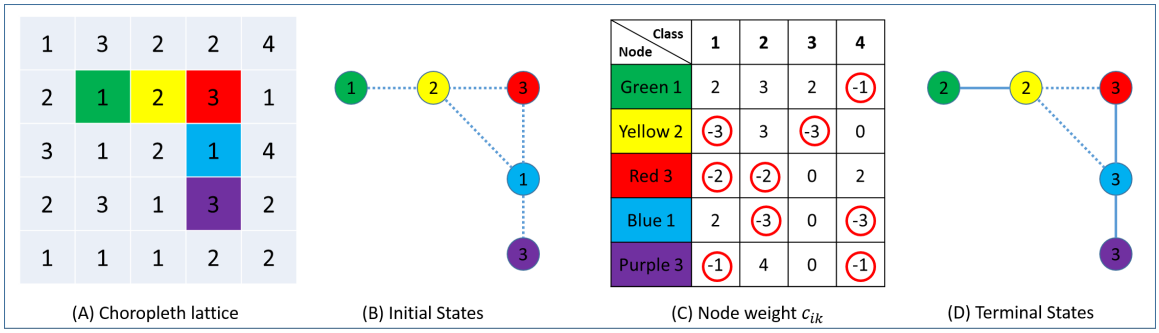


Figure 6.2: An example of adjacent changeable regions. Dashed lines represent the contiguity and solid lines represent the co-effect. Non-negative node weight indicates that class k is reachable by i. From the initial states to the terminal states, three co-effect connections have been established.

adjacent will have an entry in the spatial weights matrix $w_{ij} = 0$ making the resulting calculations independent from one another.

Once independence is established, we can identify all spatial units that fall into Cases 1-4. Then, we can consider the situation where several changeable regions are adjacent, meaning that a change of label in one region will affect the visual clustering (i.e., the value of p and q) of another changeable region. In this case, we have:

Theorem 2 *The effect of the change (EOC) only depends on the initial states and the terminal states of the changeable regions.*

Thus, the measurement of spatial association remains the same as long as the final states of those changeable regions stay the same. We generalize the effect of the

changes as:

$$EOC_{\xi} = \underbrace{\sum_{i \in \xi} (q'_i - p'_i)}_{\text{A}} + \underbrace{\frac{1}{2} \sum_{i \in \xi} \sum_{j \in \xi, j \neq i} w_{ij} (i_t \& j_t - i_s \& j_s)}_{\text{B}}, \tag{6.2a}$$

$$i_t \& j_t = \begin{cases} 1 & \text{if } i_t = j_t \\ 0 & \text{if } i_t \neq j_t \end{cases} \tag{6.2b}$$

where ξ is the set of changeable units, q'_i, p'_i are similar to q_i, p_i but exclude the other changeable units. $w_{i,j}$ is the spatial weight between spatial units i and j. i_t and j_t are the terminal states (class labels) of regions i and j respectively, and i_s and j_s are the starting labels of regions i and j respectively. Here the effect of the changes can be broken into the total separated effect caused by all of the changeable regions (Equation (6.2a) A) and the total co-effect among those changeable regions (Equation (6.2a) B). Note that the co-effect is divided by 2 because i and j are symmetric and would double the effect.

Here we can maximize EOC in Equation (6.2a) to determine the set of class labels that will create the largest visual clustering in the map. This problem can be solved by maximizing the modified Moran's I in the terminal label state of a unit. Here, we note that this may not be a desirable effect as this could introduce spurious patterns into the map; however, elements near classification boundaries need to be inspect and the EOC can be used as a metric for defining which elements could have the largest potential change on the visual output. First, it is assumed that there exists a group of contiguous spatial units which lie near classification boundaries (Figure 6.2(A)). Each unit can be altered to a certain class with a known weight. The weight is set to be the number of neighboring units that share the same class label. In practice, for the class that a unit i cannot change to, the weight is set to $-\sum_{j \in \xi, j \neq i} w_{i,j}$ (see the red circle in Figure 6.2(C)). By setting the weight to $-\sum_{j \in \xi, j \neq i} w_{i,j}$, we neutralize

the possible co-effects and guarantee that a unit cannot change into an unreachable label. If the two adjacent units have the same class label, an edge will be established with a given weight. For simplicity, the weight of the edge is unified to 1 when the spatial weight $w_{i,j}$ is 1. Finally, this can be formulated as a maximization problem where the nodes need to be labeled such that the overall weight of the nodes and edges is maximized. This can be further defined as an integer linear programming (ILP) problem. Given a graph $G = (V, E)$ with n nodes and each node has m choices of class labels, we introduce binary variables $x_{ik} (i = 1, \ldots, n, \text{ and } k = 1, \ldots, m)$ to indicate whether node i has been labeled as class k. The weights $c_{ik} \in \mathbb{R}$ are given for each x_{ik}, and variables $y_e, e \in E$ indicate whether edge e is valid based on if its two nodes have been labeled in the same class (Figure 6.2(D)). The resulting ILP can be formulated as:

$$\max \quad \sum_{i=1}^{n} \sum_{k=1}^{m} c_{ik} x_{ik} + \sum_{e \in E} y_e \tag{6.3a}$$

$$\text{s.t.} \quad \sum_{k=1}^{m} x_{ik} = 1 \qquad i = 1, \ldots, n \tag{6.3b}$$

$$2y_e - x_{ik} - x_{jk} \leq 0 \qquad e = (i, j) \in E, k = 1, \ldots, m \tag{6.3c}$$

$$x_{ik} + x_{jk} - y_e \leq 1 \qquad e = (i, j) \in E, k = 1, \ldots, m \tag{6.3d}$$

$$0 \leq x_{ik}, y_e \leq 1 \tag{6.3e}$$

$$x_{ik}, y_e \in \mathbb{Z}\,. \tag{6.3f}$$

Here Equation (6.3c) constrains two nodes of a valid edge to be in the same class and Equation (6.3d) constrains an invalid edge to not have two nodes in the same class. By solving this ILP we can identify the terminal states which maximize the Moran's I, the same formulation can also be used to minimize the Moran's I. The problem of finding the maximum possible value is similar to the Maximum Edge-Weighted Clique Problem (MEWCP) [201] which is a known NP-Hard problem [202].

Therefore, the problem of calculating the maximum EOC is also an NP-Hard problem (i.e., there is no general solution that can find the optimized value in polynomial time). Efficient algorithms for the MEWCP, such as heuristics approximation, may be modified and applied to solve this ILP. However, in practice, we find the number of adjacent nodes and the number of class choices are relatively small (traditional choropleth map design rules of thumb limit the number of classes to be less than 9). Thus, we implement a brute force solution to compute all possible values in our framework. During this computation, our framework stores the configuration of the class labels that would maximize or minimize the current spatial association. Figure 6.3 shows a multi-dimensional classification of demographic data in the United States and we highlight county boundaries based on their correspondence to the cases of Figure 6.1 for illustrative purpose.

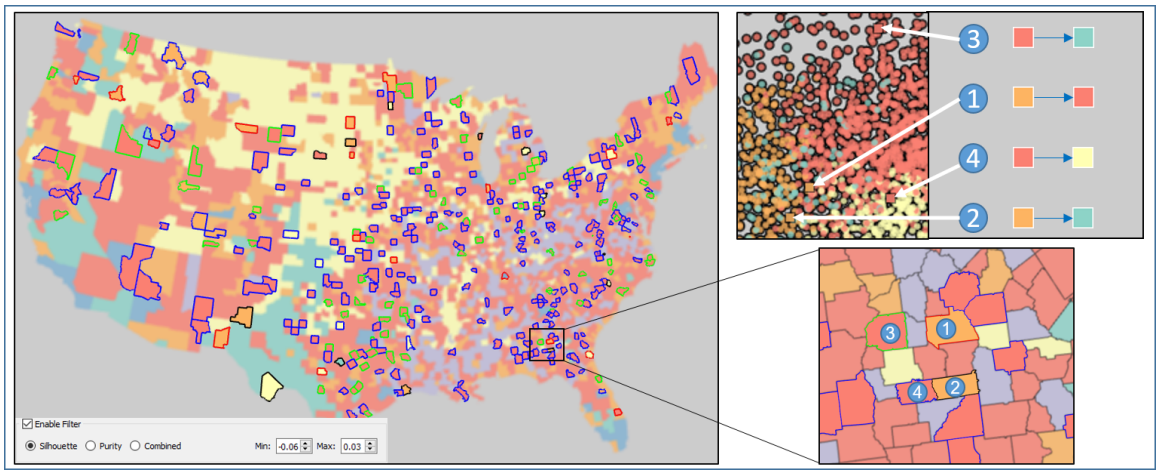


Figure 6.3: A k-means classification of US census variables illustrates boundary elements and their corresponding cases from Figure 6.1. Here the Red outline represents Case 1, the Black outline represents Case 2, the Green outline represents Case 3, and the Blue outline represents Case 4.

6.2 Summarizing the Visual Impact

While our proposed metric for quantifying the impact of visual change takes into account class labels, perceptual studies have also shown that the size of the map units is a primary driver behind the patterns that users observe. As noted by Haklay [203], a thematic map created using spatial units that vary in shape and size leads the user into thinking that the larger areas are more significant because they have a bigger visual impact than the smaller areas. Seonggook [204] proposed the concept of gross change detection and verified that different spatial distributions between two adjacent choropleth maps may lead users to under- or over-estimate the gross change in the map, which implied that the spatial distribution of change should be considered. As such, the size of the region should be considered when quantifying the visual impact of the label changes. Goldsberry and Battersby [28] introduced the magnitude of change (MOC) to quantify the graphical change between choropleth map pairs for animated choropleth maps. MOC is applicable to both object-oriented and pixel-based measures, and we extend our EOC measure to consider the size of the map element with a final metric for quantifying the impact of boundary effects on the visual spatial association in choropleth maps. The metric is a simple multiplication to derive the visual impact of changes (VIOC) and is defined as:

$$VIOC_{\xi} = \sum_{i \in \xi} \left(\frac{s_i EOC_i}{S} \right), \tag{6.4}$$

where s_i is the area size of the ith region (in pixels) and S is the overall area size of the map (in pixels). This accounts for the proportional physical change of the choropleth map under different resolutions.

Once these metrics are defined, we can now identify units on the map that could potentially be modified to change the visual appearance of spatial association. While there are methods for specifically identifying statistically significant spatial associa-

tions on a map, the majority of choropleth maps are presented with no underlying analysis of spatial association. Instead, they are presented in the wild and left solely for visual interpretation. By being able to quantify potentially spurious elements on a map, new designs could be considered where the elements could be blurred, highlighted or relabeled to another separate class in order to try and insure that patterns being seen are what was intended by the map designer (of course we recognize that the intent of the designer could have been to mislead). Thus, our method can be summarized into the following steps:

1. Choose a classification method and label the dataset of interest
2. Calculate the silhouette value for all elements in the dataset
3. For all elements whose silhouette value is within a user defined range τ calculate the EOC/ VIOC
4. Render the classified choropleth map and visually highlight all units with a VIOC value within a user defined range γ

After the map is rendered, the designer can inspect the marked units, create a map that will minimize or maximize the EOC/ VIOC, manually change units near classification boundaries to obtain the desired rendering effect, or embed the EOC/ VIOC measures as uncertainty information in the map design.

Chapter 7

CASE STUDIES

Clustering analysis in geospace requires lots of domain knowledge about the data to generate valid case studies. For instance, demographics classification is a process that not only involves geographic information but also needs other information, such as economic and social background, to label the results. To evaluate the proposed framework, case studies using several datasets including conflict in Africa, crime statistical data, economic development, and demographics data are explored.

7.1 Conflicts in West Africa

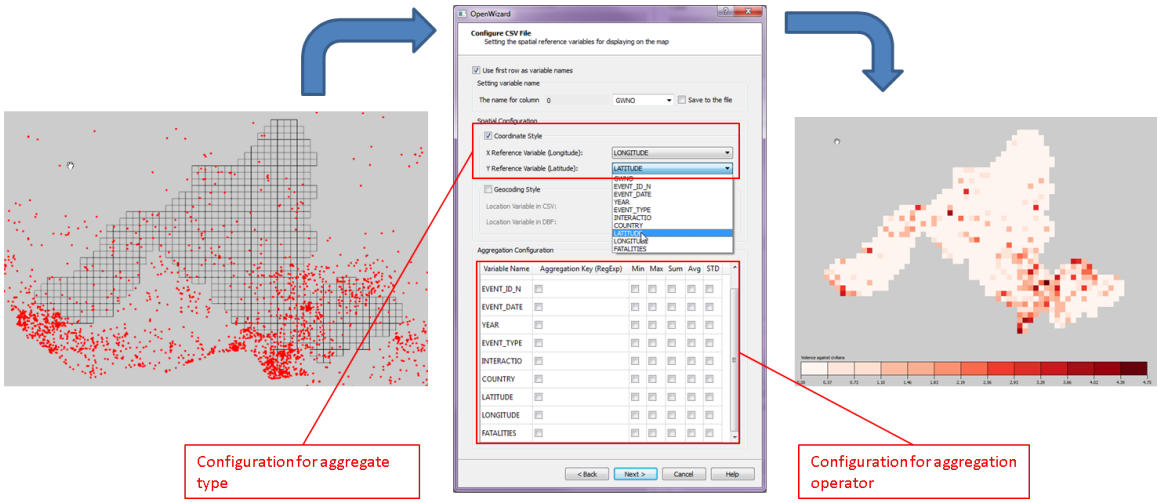


Figure 7.1: Aggregation of conflicts in West Africa. Left part is the conflict events in the whole Africa. Middle part is the Open Wizard with all the wrangling configurations. Right part shows the choropleth map after aggregating the conflict events to the grid cells.

The first case study focuses on conflicts in West Africa. The base dataset is a .shp file of 828 grid cells of 0.5 X 0.5 degrees derived from PRIOGRID [205]. This dataset only contains 3 variables which are the grid cell ID, country name, and administrative area name. The conflict data comes from another .csv file that has 8 attributes (country code, country name, conflict ID, event date, conflict type, longitude, latitude, and fatalities). To combine these two datasets, the users applied our data wrangling procedure by loading both files into the open wizard and checking "count" as the aggregation operator for the event ID (Figure 7.1). After the aggregation procedure, users could directly visualize the number of conflicts in each grid cell of West Africa in the choropleth map.

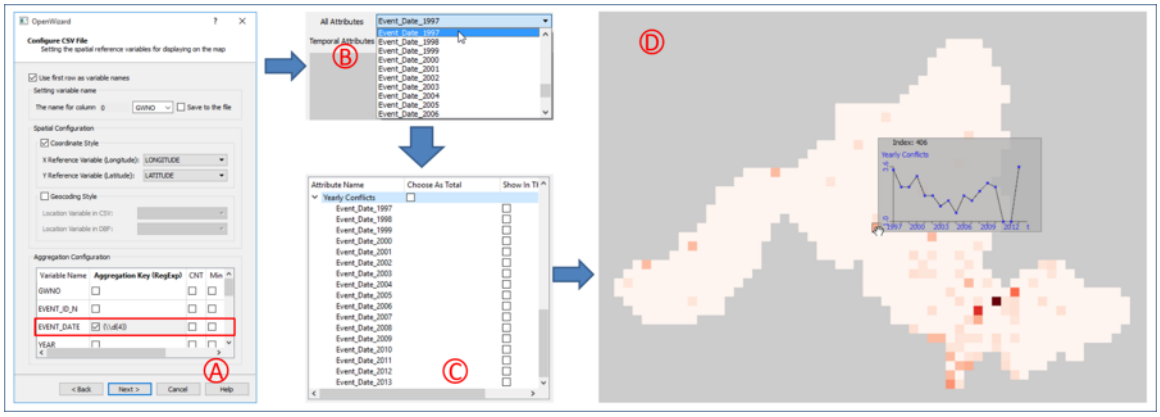


Figure 7.2: An example of using regular expression with aggregation key. The process from A to D demonstrates how users can achieve temporal variable from the spatial aggregation.

In this example, there are 10 conflict types distinguished by categorical numbers, so users could also compute the statistics of conflicts based on the conflict type. Users checked the aggregation key option for the conflict type and aggregated the data again. The framework automatically detected the types of conflict events and generated counts for each of the conflict types. Furthermore, users were interested

in investigating the yearly temporal aspect of the conflicts. To their knowledge, the event date is in a format of "MM/DD/YYYY". Thus, by entering the regular expression "`(\\d{4})`" in the text field of aggregation key under the event date attribute (Figure 7.2(A)), the conflicts will be aggregated in a yearly manner. After the aggregation, the temporal configuration page of the Open Wizard has been used to group variables temporally (Figure 7.2(C)). In this way, users can directly visualize the time series of conflicts on the map (Figure 7.2(D)). Feedback from users showed that the wrangling process in this framework greatly reduces the time and energy cost of data preprocessing. Without such wrangling process, typically it could take several days to code scripts and generate the same results.

7.2 Crime Estimates in the United States

In this example, crime data from the U.S. Department of Justice are used. This is an annual dataset that contains 19 numerical variables for 49 states of the mainland US spanning from 1990 to 2012. Here, users are interested in exploring property crime rates. They begin by exploring various crime thresholds in the dataset using the interactive histogram widget, Figure 8. By dragging the threshold from the lowest property crime rate to the highest property crime rate they can observe overall trends in the US. In Figure 7.3(1), a few states with a decreasing trend at this level have been filtered out by setting the threshold value low. This indicates that these states have some of the lowest property crime rates in the country. As the user moves the slider Figure 7.3(1-3), they begin seeing patterns of decreasing trends and oscillations; however, no increasing trend (highlighted by green) is evident. This indicates that there is no consistent rise in property crime rates for any of the study regions; the majority of states have crime rates that either decrease over time or oscillate above and below the threshold crime rate for. The users note that they think of this as

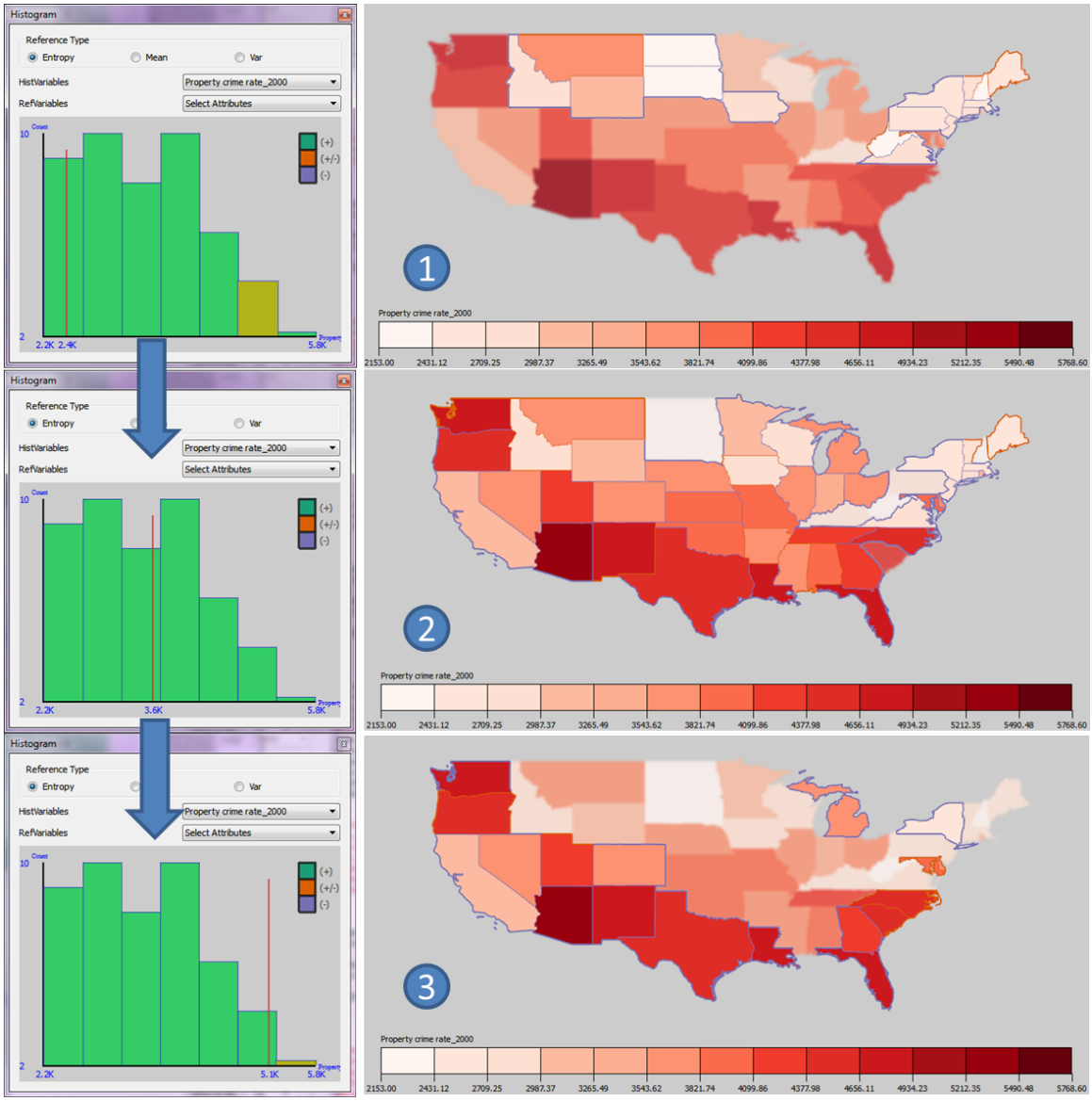


Figure 7.3: Using the threshold widget as an overview mechanism for exploratory data analysis. By changing the threshold, users can explore the relationship between crime levels and state wide trends.

a way of looking at levels of support or resistance for criminal activities and would make use of this information by linking it to other policy datasets for future analysis.

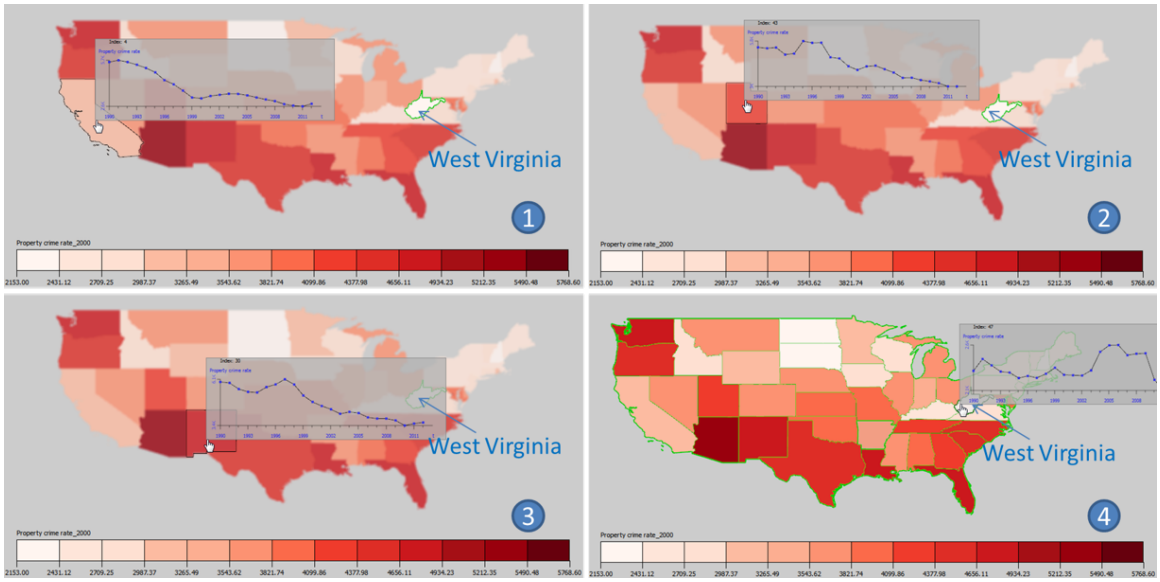


Figure 7.4: Users explore temporal similarity through a value based filter. (1-3) show the brushing result of less than 20% similarity in three different states respectively which all highlight West Virginia. (4) shows the brushing result of less than 20% similarity in West Virginia which highlighted all other states.

Next, the users wanted to explore regional trends using similarity brushing. As they know most states have a decreasing trend in property crime rates, thus they would like to see if there are any outliers. They select a Sequential Normalized Distance metric with a similarity threshold less than 20% to highlight states that are least similar to one another. When brushing various states, West Virginia was highlighted in the result, Figure 7.4(1-3). Then, the users brushed West Virginia and saw that all other states were highlighted (Figure 7.4(4)) indicating that property crime rate in West Virginia is very different from other states. To further investigate the relative similarity between West Virginia and other states, the users choose the rank option with the threshold of larger than 80% which is the top 9 most similar trends out of 49 trends. The result (Figure 7.5(a)) highlights 9 states on the map that are supposed to be relatively similar to West Virginia in terms of property crime

rate trends. However, by accessing the temporal trend multiples (Figure 7.5(b)), the users can tell how relative these similarities are. This result highlights the value of our visualization tools in revealing deceptively similar trends in similarity metrics.

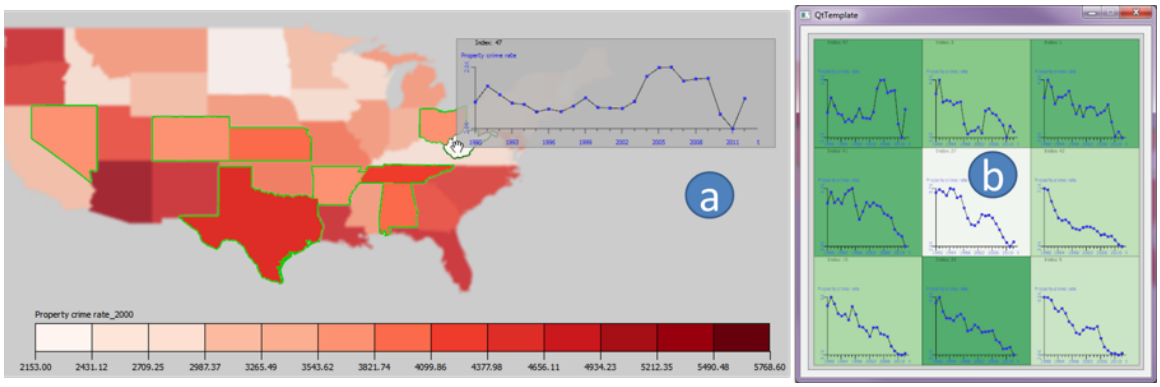


Figure 7.5: Users explore temporal similarity for West Virginia through a rank based filter. (a) 9 states that have relatively similar trends with West Virginia (b) Temporal trend multiples help users visually inspect how similar West Virginia's temporal trend is to other highlighted states' trends.

7.3 Indices of Industrial Diversity

In this study, data were obtained from the U.S. Census Bureau's County Business Patterns (`http://www.census.gov/econ/cbp/`) database, which contains detailed industrial information about establishments for all U.S. counties dating back to 1986. The variables of industrial diversity are measured by the number of establishments, which are subdivided by two digit North American Industrial Classification System (NAICS [206]) industry. There are 17 two-digit NAICS industries, 11 years from 1998 to 2009 for each of those industries in this study and 3,106 counties in the continental United States for which data are available. The data file is provided in .dbf table format corresponding to the shapefile of counties in US. This time series of 1998-2009 cover the most recent economic crisis, in which the U.S. entered a severe recession

following the bursting of the housing bubble in mid-2007. To analyze spatio-temporal trends in specific industries hit hard by the recession, the user identifies several major metropolitan areas that have a high concentration of establishments in three industries: Construction (N23) and Real Estate (N53), and Entertainment (N71). Each of these industries is anticipated to have been hit hard by the recession because Construction and Real Estate are directly related to the housing industry while Entertainment is likely to have been impacted as people cut back on non-essential expenses due to job losses. To facilitate this, the user first examines the multivariate similarity between Construction, Real Estate and Entertainment (Figure 7.6(a)). According to their expert knowledge, the user brushes over Maricopa County (i.e., the Phoenix metropolitan) in Arizona, that was known to be one of the hardest hit regions of the country during the crisis [207]. The result returned the major metropolitan areas that are most similar to Maricopa County including Los Angeles, Seattle, Houston, New York, etc.

After this process, the user wanted to know whether these major metropolitan areas have similar temporal behaviors with respect to these three variables (Construction, Real Estate, and Entertainment) as well. To do this, the user set up three tooltip temporal filters corresponding to Construction, Real Estate and Entertainment respectively using Sequential Normalized Euclidean distance metric with a threshold of 80%. Figure 7.6(c, d, e) shows the counties with temporal trends similar to each of the three underlying variables. The user quickly notes the major metropolitan areas (e.g., Las Vegas, Chicago, Boston, and New York) across the country with similar trends across each of these three variables. However, the Houston metropolitan area, of which Harris County is a part, does not exhibit a similar trend. This matches previous research indicating that the housing crisis was less severe in Texas

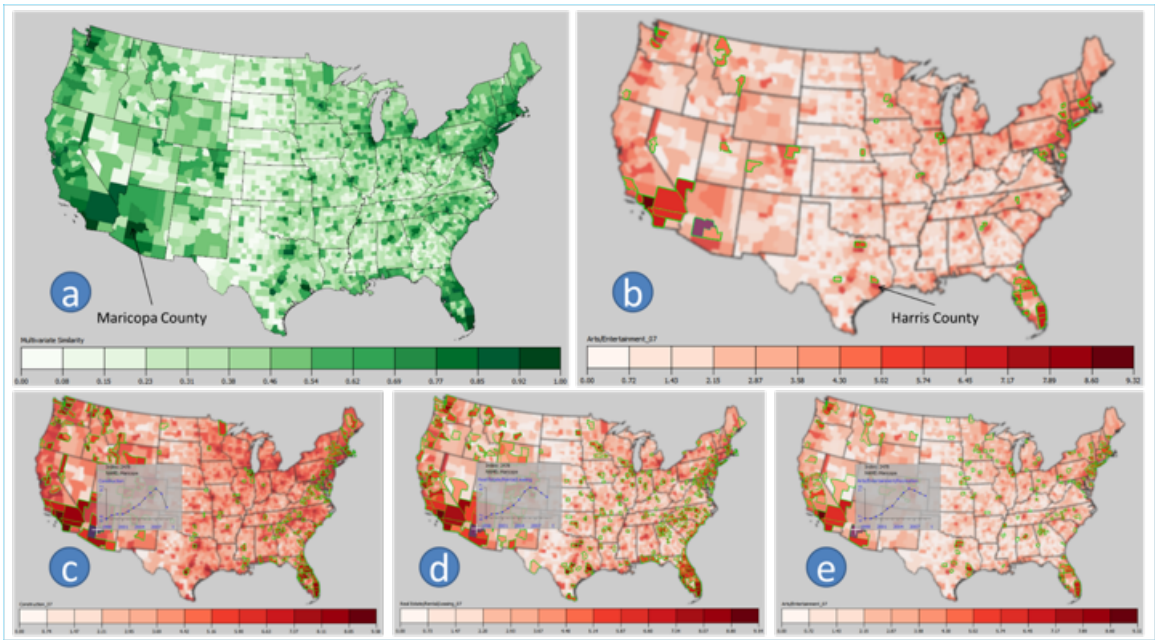


Figure 7.6: Exploration of the indices of industrial diversity dataset using analytical brushing. In part (a), the users applied multivariate distance mapping to find the similar units in multivariate space with respect to Maricopa County. In part (c, d, e), the users would like to find the temporal similarity using analytical brushing. They employed a Sequential Normalized Euclidean distance metric with a threshold of 80% similarity with respect to Maricopa County, AZ. From left to right: Similarity with Construction (N23), Similarity with Real Estate (N53), and Similarity with Entertainment (N71). After exploring about the similarity for each criteria, they explored the result of an AND operation using the previous three brushes, results of which are shown in (b).

than in other parts of the country [208]. As such, Harris County largely avoided the great recession due to its state's liberal and market oriented land use policies [208].

The user was also interested in learning about whether there are counties that have an upward trend rather than just a delayed or less severe crash in Construction. The user started with the time series from Harris County. In order to focus just on

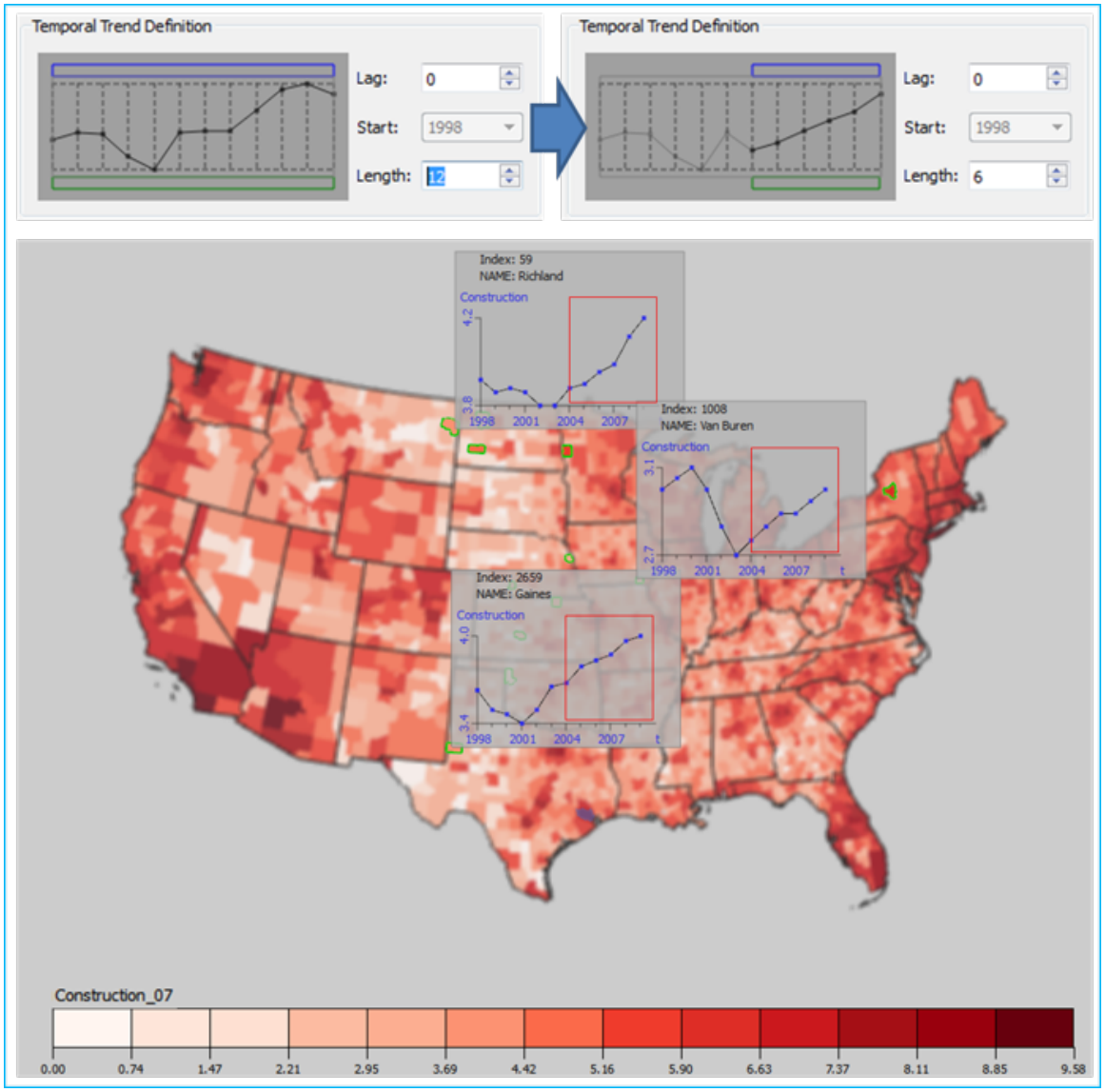


Figure 7.7: User defined trend exploration. The map at the bottom shows the result of similarity brushing for the user defined trend after applying a Sequential Normalized Distance with threshold larger than 80%.

the period of the recession, they adjusted the length of analysis to cover only the last six years (Figure 7.7). The user applied the temporal similarity brushing again but used a higher similarity setting of 90%. The results showed several counties that are not metropolitan areas having an ascending trend after the housing bubble

burst. Furthermore, the user investigated one of them: Richland County. Due to the high-wage employment growth in Richland from oilfield development, the population growth rate has surpassed the regional average in every year since 2005 [209].

Feedback from the domain expert on this paired analysis gave positive feedback. The expert noted that such analyses would be difficult to do in other systems, particularly, the multivariate combination to explore similarity over space with respect to multiple variables. While in the univariate case, a parallel coordinate view could provide a temporal trend analysis of a single variable, finding similar trends across multivariate data would be extremely difficult. Furthermore, within this problem domain, there are over 3000 counties in the United States, there are 19 temporal measures of industrial composition per county, and each of these measures is used to compute a location quotient for each county as well as a series of index values describing the county's overall specialization. Thus, given the large number of counties and time series, it is intractable to replay every temporal animation of every single variable to look for changes over time. Key features the expert used was the temporal threshold widget for defining thresholds of interest, and the temporal definition widget where the expert could draw various trend lines and then combine them in the logic widget.

7.4 Impact of Geographical Variations on clustering

In the following case studies, we used a demographics data set containing quick facts about counties in the continental United States from the US Census Bureau (`http://quickfacts.census.gov/qfd/download_data.html`). There are 3106 counties in this dataset and 52 demographic variables. The counties are distinguished by their FIPS (Federal Information Processing Standards) number, and the variables by their

mnemonic identifier. Note that variable choices here are chosen to clearly highlight observable patterns in the data and demonstrate our framework features.

7.4.1 Discrete Spatial Extent

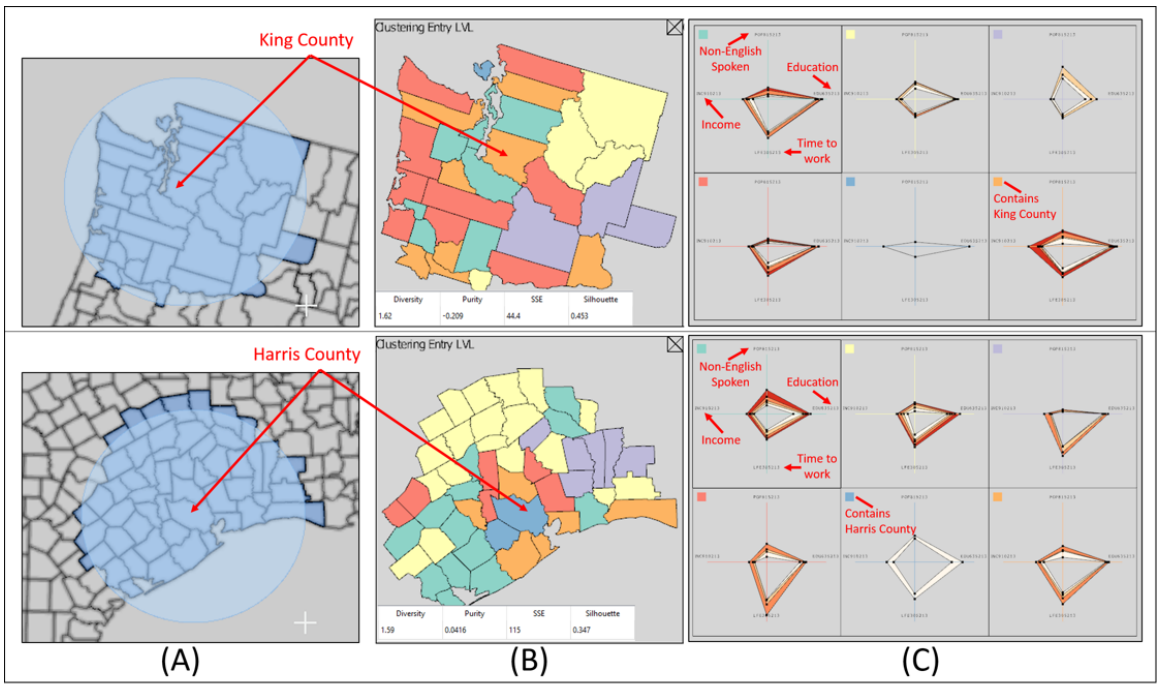


Figure 7.8: An example of exploration between clusterings in different geographical locations. The top and bottom row are results based on surroundings of King County and Harris County respectively. (A) Selections with the circle selection tool. (B) Local clustering results and its clustering statistics. (C) Rose plots for the local clusters, the variables from the top in a clock-wise manner are: percentage of other languages, percentage of education level above high school, mean time to work, and per capita income.

POP815213 (Language other than English spoken percentage), EDU635213 (Education level above high school), LFE305213 (Mean travel time to work), and INC910213 (Per capita income) are the variables of interest in this case study. Here we explore

the surroundings of King County (Seattle) and the surroundings of Harris County (Houston) to determine if they share any common patterns in clustering. We first choose the surrounding counties within the same radius from both counties using the circle selection tool (Figure 7.8(A)) and then apply hierarchical clustering using Ward's method [210] and Euclidean distance (Figure 7.8(B)). The relatively high value of the average Silhouette coefficient from King County's clustering indicates the goodness of its clustering is slightly better than Harris County's clustering under the same clustering method. From the rose plots (Figure 7.8(C)), we also notice that the characteristics of each cluster in King County's clustering are more distinguishable. As King County belongs to the orange cluster, we identify that it possesses more well-educated people, higher income, and requires more time travel to work when compared with the other clusters from the rose plots (Figure 7.8(C) Top). The difference of cluster characteristics is small from King County westward, but is large eastward as these counties have significant less travel time to work and more non-English language speakers. Harris County is similar to the situation in King County; however, the difference of cluster characteristics does not shows the similar west-east pattern. We can tell the non-English spoken percentage drops significantly towards the north but remains towards the south, which makes sense as Mexico boarders Harris County in the south.

Interestingly, the other two values of Shannon diversity and average purity from Harris County's clustering suggest that the clusters in Harris County's clustering are more spatially associated and the clusters in King County's clustering are more scattered. The spatial distribution of the clustering results in Figure 7.8(B) display the corresponding patterns. Though these two counties are both near the ocean, close to the country border, and contain major metropolitan areas, the different styles in

population composition, education level and traffic level of its surroundings counties appear to contribute to different spatial patterns.

7.4.2 Discrete Geographical Features

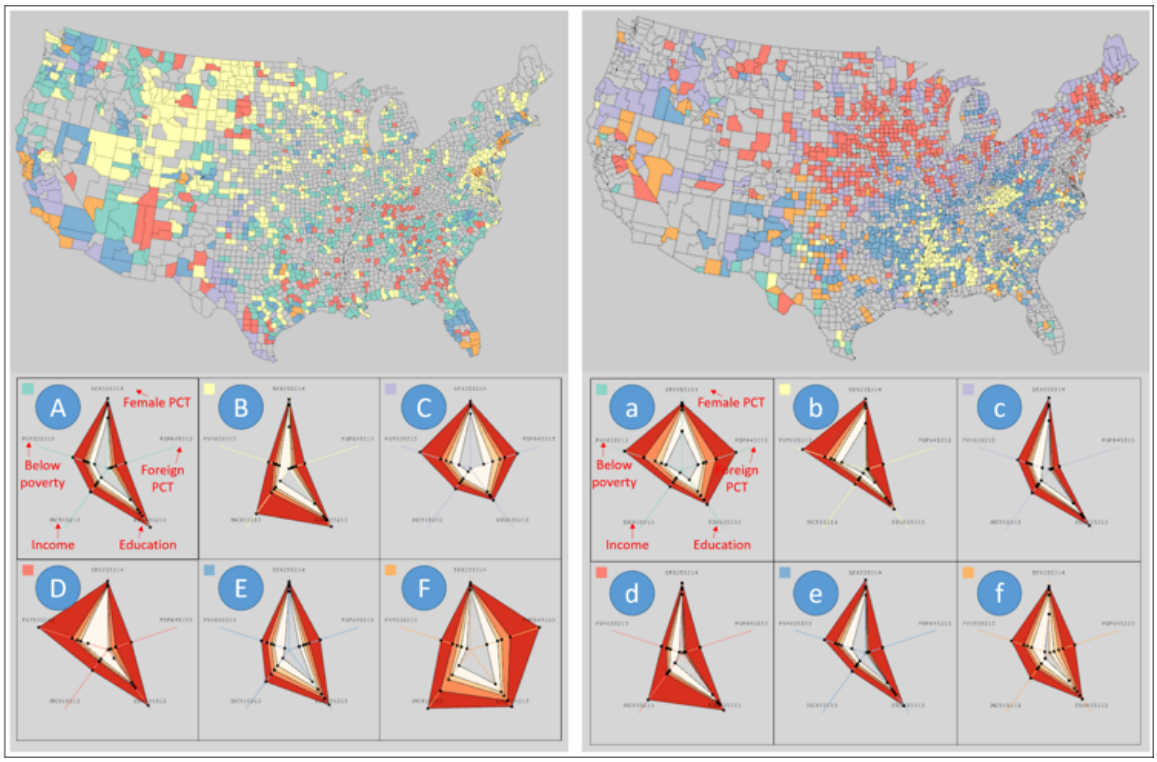


Figure 7.9: An example of partial clustering for units of different geographical feature. The left map shows the units with all positive population change and the right map shows the units with all negative population change in 2014.

We noticed that variable PST120214 representing the percent change of population in 2014 has both negative and positive values, so we explore clustering on units with positive values and negative values respectively. We first use the scatterplot (or categorical view) to select units with only positive values and then apply hierarchical clustering using Ward's method and Euclidean distance. The number of clusters chosen is 6 and the variables clustered on are SEX255213 (Percent of female in 2013),

POP645213 (Percent of foreign born persons in 2013), EDU635213 (Percent of persons with high school graduate or higher), INC910213 (Per capita money income in 2013), and PVY020213 (Percent of persons below poverty level). Then we repeat the same process for the units with only negative percent change of population. Figure 7.9 demonstrates both the clustering results for units with positive and negative population change.

After investigating the rose plot for each clustering result, we find a similar matching pattern based on the clusters' characteristics. As shown in Figure 7.9, a matching of $A \Leftrightarrow e/c$, $B \Leftrightarrow d$, $C \Leftrightarrow a$, $D \Leftrightarrow b$, $E \Leftrightarrow f$ can be easily identified from the rose plot of each cluster. This means that spatial units with positive and negative population changes do share some similar patterns with those 5 variables. We also notice that only the F cluster does not have a matching cluster in the other clustering result. Cluster F has the highest distribution in education level, income level and foreign born persons level. Also, judging from the map, we can tell that the units from cluster F are all major metropolitan areas such as Los Angeles, New York City, etc. We hypothesize that areas with high income levels, education levels and foreign born population will have positive population change, in other words, areas with negative population change usually do not have high income, education and foreign born persons level.

7.4.3 Continuous Spatial Extent

Next, we explore the scale change effects in clustering. The clustering features 6 clusters with Ward's method and Euclidean distance. It is based on the variables of age (percents of person under 5, under 18, and above 65 respectively), house living (percent of living in same house more than 1 year), and education (percent of person have bachelor's degree or higher). We first choose Cook County (the Chicago metro

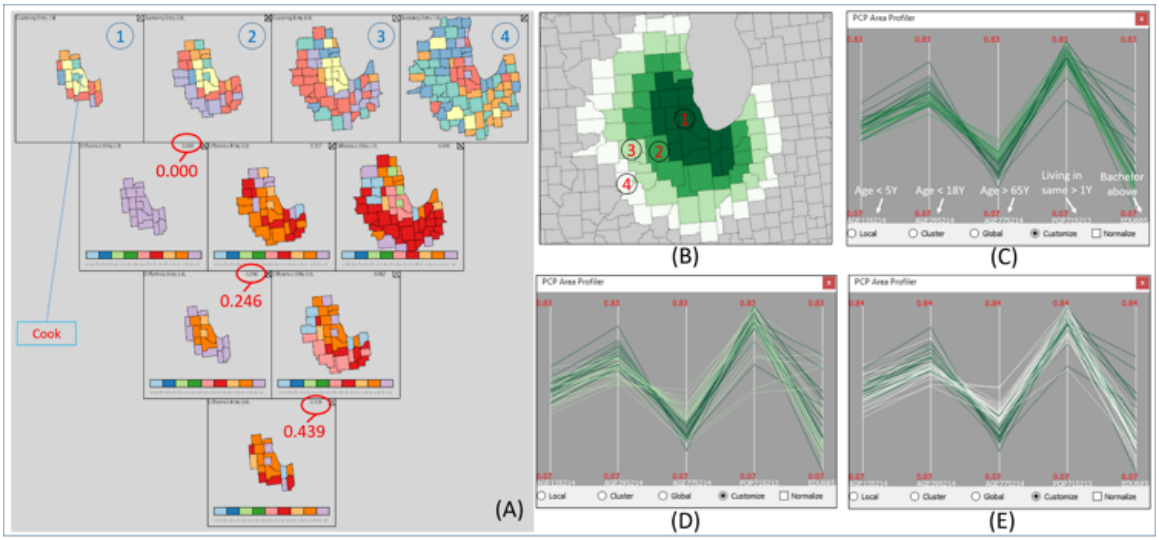


Figure 7.10: An example of scale effect on clustering around Cook County. (A) The Triple-D view of the clustering comparison regarding the scale change (B) The change of scales demonstrated in 4 colors (C) PCP area profiler for scale 1 and 2 (D) PCP area profiler for scale 1 and 3 (D) PCP area profiler for scale 1 and 4.

area). We select the start scale and end scale as shown in the Figure 7.10(A-1) and (A-4) respectively. Then the framework automatically interpolates the steps between those two scales (Figure 7.10 (A-2)(A-3)). After clustering on each of the scales and applying the comparison in the Triple-D view, we find out that the comparison metrics which stands for dissimilarity are slowly increasing as the scale changes (Scores circled in red in Figure 7.10(A)). Thus we label the change of scales in different colors (Figure 7.10(B)) and visualize the difference between them with PCP area profiler (Figure 7.10(C)(D)(E)). As shown in the PCP area profiler, Cook County and its neighboring counties have a higher measurement in the % of population under 18 and in the education variable (Darker Green lines as scale 1 in Figure 7.10(C)). When the scale expands outward to the next contiguous set of neighbors, the outer counties have higher percent of elderly and the education level goes down (Yellow lines as

scale 1 in Figure 7.10(C)). As the characteristics of these two area profilers are quite distinguishable, it means that the scale change from A-1 to A-2 (B-2) may have less effects on the clustering of A-1. That also explains the indiscrimination of the comparison between the scale A-1 and A-2's clustering results. When the spatial extent increases, more units that are similar to the units in scale A-1 have been induced (Overlapping purple lines in Figure 7.10(D)), and that interferes with the clustering results. Note that clustering results here are further confounded due to variables being dependent proportions of the population.

7.4.4 Continuous Geographical Resolution

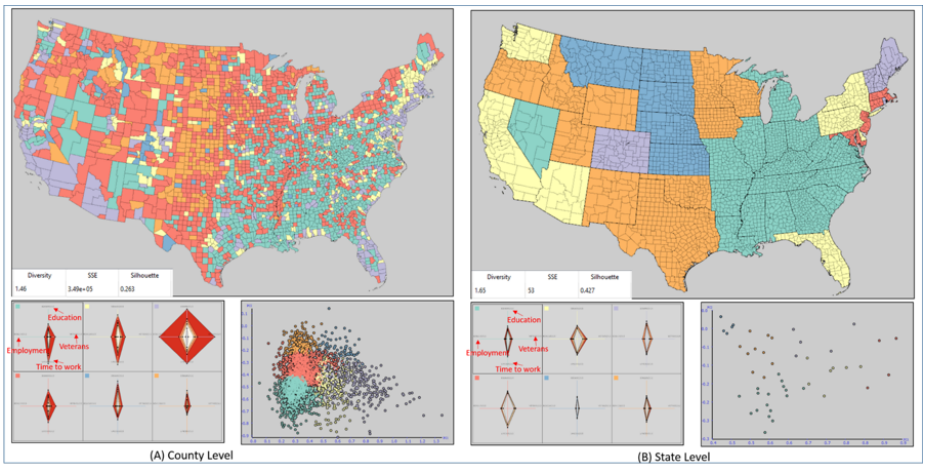


Figure 7.11: An example of clustering results under different geographical resolutions. (A) Clustering result, corresponding rose plot and PCA scatterplot in county level, (B) Clustering result, corresponding rose plot and PCA scatterplot under state level. Both use the same hierarchical clustering with 6 clusters.

As shown in Figure 7.11, we applied the hierarchical clustering of 6 clusters on the county level and state level respectively across the mainland US. The clustering uses the Education level (EDU685213), Veterans (VET605213), Mean travel time to

work (LFE305213), and Private nonfarm employment (BZA110213). We can tell that county level rose plots (Figure 7.11(a)) have more variance for each of the clusters. The PCA scatterplot also shows more overlap at the county level, but clear separation in the state level. From the statistics of the two clustering results, the average Silhouette coefficient of clusters under state level is higher than under county level which indicates higher intra-cluster similarity at the state level aggregation. These observations in cooperation with the work of Sun [211] demonstrate that spatial aggregation can improve data quality.

7.5 Visual Impact of Changes in Classification Boundaries

In this section, we demonstrate by example that our proposed metrics are able to identify map elements whose labels lie near classification boundaries and that changing these elements' classification label will impact the perceived visual spatial association.

7.5.1 Applying EOC for Visual Clustering

The first dataset used here is the Chicago crime data of 2014. There are 77 regions and 26 types of crime variables in this dataset. For classification, k-means clustering has been applied with $k = 5$ for three variables "Liquor Violation", "Sex Offense" and "Robbery". The resulting classification and the PCA scatterplot are shown in Figure 7.12.1. All spatial units are outlined to provide a geographic overview for discussion. Next, we identify spatial units that may be near the cluster boundary by setting the silhouette value to the range of -.2 to .2. Units in this range that will have an impact on the EOC are outlined on the map and in the PCA scatterplot (Figure 7.12.3). Figure 7.12.2 shows the results of minimizing the EOC to reduce

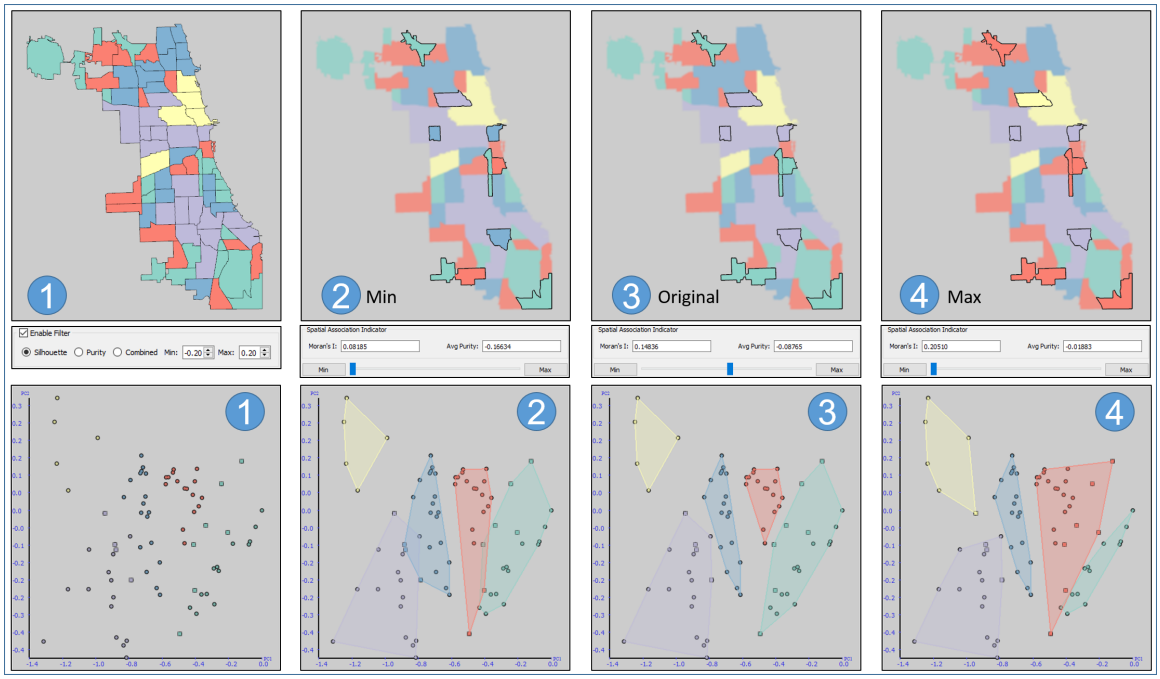


Figure 7.12: Minimizing and maximizing the EOC of elements near the classification boundary using criminal incident reports in Chicago, IL. 1 - The initial k-means classification. 2 - Minimizing the EOC, thus creating more visual heterogeneity, highlighted elements are those near the classification boundary. 3 - The initial k-means classification with only the changeable elements highlighted (provided for quick comparison purposes only and is the same as 1). 4 - Maximizing the EOC, thus creating more visual spatial clustering.

the spatial clustering that is visually observed, and Figure 7.12.4 shows the results of maximizing the EOC to increase the spatial clustering that is visually observed.

To further summarize, our modified Moran's I in Figure 7.12.3 (the initial k-means clustering) is .14836. By shifting the cluster labels as in Figure 7.12.2, Moran's I can be reduced to .08185 and more dispersion in the regions is seen. For example, the large purple region in the middle is dispersed as the unit on the Purple/Blue boundary shifts to Blue. Such an effect may be desirable in map design as this may help eliminate

the potential for users to identify spurious patterns if the result of the visualization is designed to be as disperse as warranted by the data. Note the shift of the convex hulls in the scatterplot as well when the units are re-labeled. In Figure 7.12.3, we see the Purple-Blue border now overlaps as does the Red-Teal. Similarly, in Figure 7.12.4, Moran's I can be increased to .20510 and we see larger red regions form in the North and South. The Red-Teal border now overlaps in the scatterplot as well.

The most interesting boundary in the PCA projection is the Teal-Red boundary. In the Teal group, measures of the three crimes are all quite low; however, in the Red group, the data is clustered around mid-level rates. Rates are normalized by the total count of crimes, and in the Red group, liquor violations and sex offense have normalized values ranging from .19 to .43 and .21 to .47 respectively. In the Teal group these rates are 0 to .19 and 0 to .047 respectively with robbery rates in both groups being less than .21. Thus, if one were to provide a label to the Teal cluster, it could reasonably called the "low risk group" and red could be a "mid-to-high risk group". What we see in Figure 7.12 is that there are regions in the North and South of Chicago with a Teal unit surrounded by Red. When we maximize the EOC, the Red clusters become visually larger indicating more areas in the "mid-to-high risk group." Given that the units that were changed are near the classification border, the change from Teal to Red could be warranted, and the designer's goal could be to show that crime is a problem in Chicago and larger visual clusters could help sell that point. Again, the goal of this thesis is not on the ethical implication of such design choices, but the focus is on the fact that elements near classification borders may need to be identified to capture a holistic picture of the multivariate classification scheme.

To further demonstrate the impact of minimizing or maximizing the EOC we explore demographics data for counties in the Western United States. The data is obtained from the US Census Bureau (`http://quickfacts.census.gov/qfd/`

`download_data.html`). We perform k-means clustering with $k = 6$ on three variables, PVY020213 - percent of population below poverty, LFE305213 - mean travel time to work, and EDU685213 - percent of population with education higher than a bachelor's degree. Results of the clustering are shown in Figure 7.13. Figure 7.13.1 highlights all the units that can impact the EOC calculation and Figure 7.13.3 is the same figure, just no highlighting in order to demonstrate how the visual clustering might be observed. Figure 7.13.2 is the result of minimizing the EOC. When the EOC are minimized (resulting in less visual spatial clustering), changes can be identified in the middle of Washington (the yellow region that was previously there has been dispersed), the middle of Colorado, and the South-West corner of New Mexico (among others). Figure 7.13.4 is the result of maximizing the EOC. Here, more visual clustering can be observed particularly in Oregon where a large segment of the state is shifted to the same cluster label.

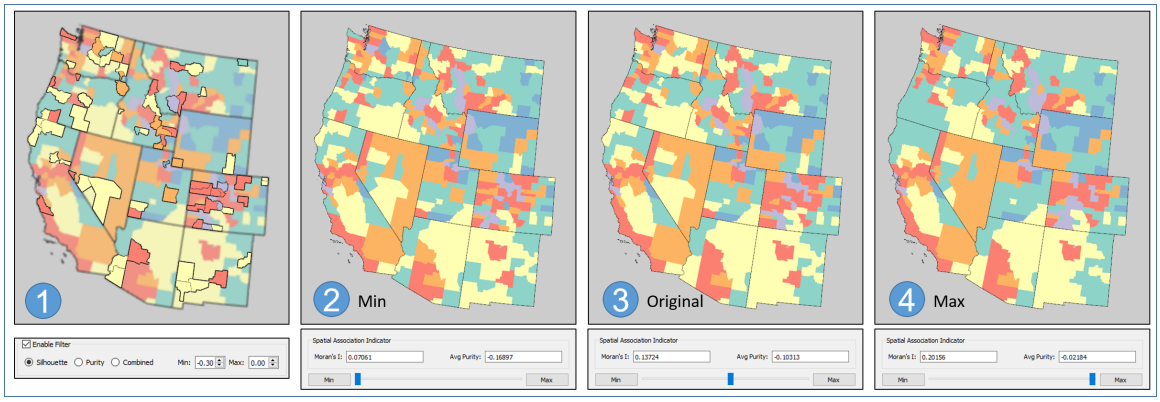


Figure 7.13: Minimizing and maximizing the EOC of elements near the classification boundary using US census data from the western United States. 1 - The initial k-means classification. 2- Minimizing the EOC, thus creating more visual heterogeneity. 3 - The initial k-means classification (provided for quick comparison purposes only and is the same as 1). 4 - Maximizing the EOC, thus creating more visual spatial clustering.

7.5.2 Combining EOC and VIOC

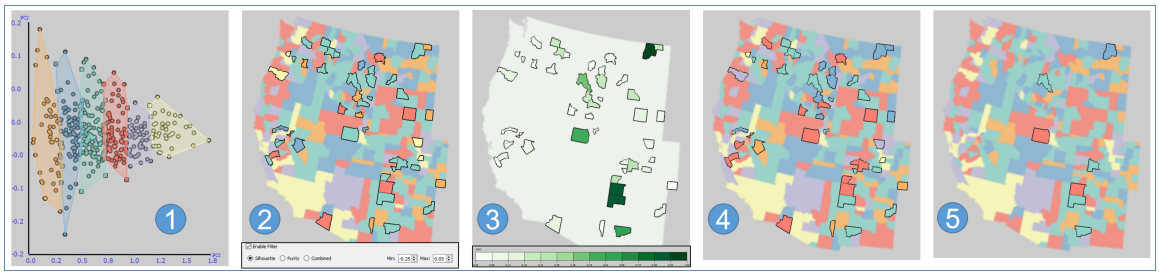


Figure 7.14: Maximizing the EOC based on the VIOC near the classification boundary using US indices of industrial diversity from the western United States. 1 - The PCA scatterplot for the initial k-means classification. 2 - A choropleth map of the k-means classification, highlighted elements are those near the classification boundary. 3 - The VIOC measure of elements near the classification boundary. Darker elements will have a larger visual impact if their label changes. 4 - Maximizing the EOC of all units near the classification boundary. 5 - Maximizing the EOC of all units with VIOC in between .46 and 1.

In Figure 7.12 and Figure 7.13, what becomes obvious is that the size of the spatial units plays a large role in the visual output. This is completely expected as documented in the related work [203]. Thus, while Figure 7.12 and Figure 7.13 focus on highlighting all units that can change with the EOC measure, the proposed VIOC measure can provide information about which units can be changed and, if changed, will have the largest visual impact. In this example, we explore measures of industrial diversity in the Western United States. These measures represent the relative concentration of industries for a given spatial unit of interest at a particular point in time. Figure 7.14 shows the result of applying k-means clustering ($k = 6$) to the indices of healthcare (N62), finance and insurance (N52), and professional and science services (N54). We use a silhouette value range of -.25 to .05 and Figure 7.14.1

and Figure 7.14.2 show the result of the classification with the units with labels on a classification boundary highlighted. Note that while other units may be initially highlighted with the silhouette coefficient, by using the EOC case criteria we reduce the highlights to only those units that will have a visual impact on the map. Note that there are approximately 30 counties highlighted on the map and we want to explore which units will have the most visual impact. We use a sequential color scheme to shade the highlighted units based on their VIOC measurement, which is directly proportional to the percent of the screen space that the spatial unit occupies. The result is shown in Figure 7.14.3. As expected, the larger the county, the darker the highlighting. The reason we show this is that by simply applying silhouette filtering and EOC metrics to highlight the boundary regions, many units will be selected. If we only want to focus on the most visually salient units, the units could be further filtered based on their VIOC values. In Figure 7.14.5, we set a VIOC range from .46 to 1 leaving only 7 of the initial 30 counties highlighted. We then modify the labels to maximize EOC. Filtering by VIOC can be thought of as another tool for the map designers toolbox in which they can consider modifications to class labels and boundaries.

7.5.3 Relabeling versus Boundary Modification

Throughout the previous examples, we have primarily discussed the impact of relabeling element that are on classification boundaries; however, simply relabeling an element may not be the most appropriate means of adjusting the classification. In multivariate schemes, such as k-means, recent work has focused on incorporating user feedback into the classification model [154]. Thus, if a user changes a label, the classification model will update the weights and reassign the classification boundaries. Recent work on this topic was discussed in section 2.6, and this concept is extended

to incorporate a modifications for flexible direct manipulation. To apply a model manipulation scheme, first, we identify elements to relabel using the silhouette range τ. We can then maximize (or minimize) the EOC which forces the relabeling of elements. This relabeling will automatically update the weights of the k-means clustering, and a new classification based on the updated weights will be generated. This result is shown in Figure 7.15. Here, we revisit the data and classification scheme applied in Figure 7.12 (the Chicago crime data).

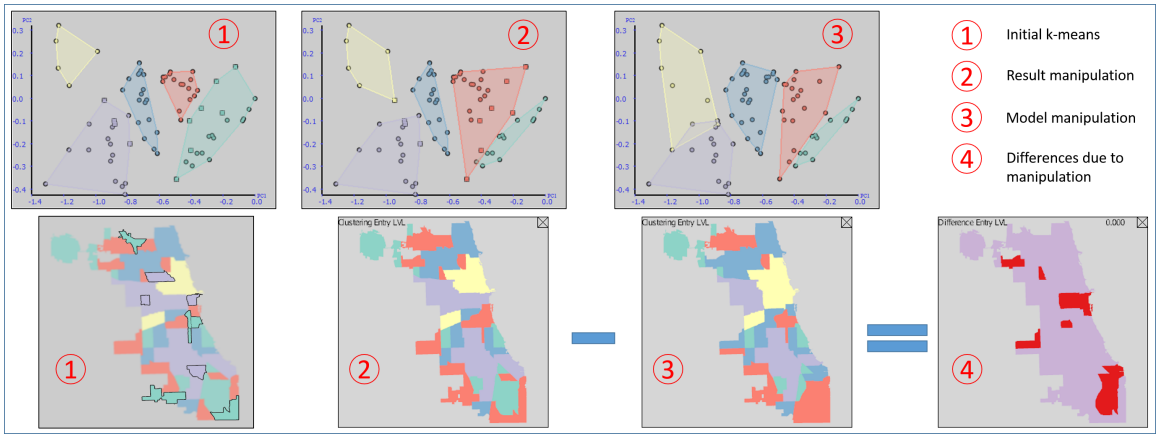


Figure 7.15: The effects of model manipulation on choropleth map classification. 1 - A k-means classification of criminal incident reports in Chicago, IL. 2 - Maximizing the EOC through result manipulation (i.e., changing the unit label does not effect the k-means weight). 3 - Maximizing the EOC through model manipulation (i.e., changing the unit label updates the k-means weights). 4 - The difference between result manipulation and model manipulation. Note that units that were not originally highlighted as being near the classification boundary are now reclassified due to the updated weights used in k-means.

Figure 7.15.1 and Figure 7.15.2 are the same k-means and maximized EOC results from Figure 7.12.1 and Figure 7.12.4 respectively. What is interesting is that by changing the k-means weights, the classification boundaries shift and units that

were not marked as boundary candidates are now subsumed by a new class. In Figure 7.15.3, the same units that were relabeled in Figure 7.15.2 are relabeled in Figure 7.15.3. This causes the weights in the k-means clustering to update, and then a new k-means classification is performed, resulting in the map classification of Figure 7.15.3. If we take a difference between Figure 7.15.2 and Figure 7.15.3, we can see what other units were shifted as a result of updating the weights of the k-means classification (Figure 7.15.4), and we notice that an even larger amount of visual spatial clustering can be seen in Figure 7.15.3 than in Figure 7.15.2.

Chapter 8

CONCLUSION AND FUTURE WORK

The case studies provided in this thesis demonstrate the utility of the wrangling process and the analytical brushing in enabling researchers to quickly form hypotheses and explore spatial and temporal trends within their data in novel ways. The wrangling procedure greatly supports the exploration process by reducing the time and energy cost required for data preprocessing. For temporal exploration, instead of relying on black box methods such as self-organizing maps [23], this framework allows users to create and shape their own logic functions to define spatiotemporal regions of interest and project these regions onto the map. This framework does not provide a direct comparison of tasks between animation, small multiples and analytical brushing due to the fact that the analytical brushing directly lends itself to answering questions such as Which regions have similar trajectories over time? By utilizing analytical brushing solutions, the cognitive burden of mentally integrating small multiples or animation can be reduced. In this way, this framework helps to tighten the visual analytics pipeline by providing brushes where the analytic algorithms are directly linked to the brush tip. Other solutions, such as animation and linked views, require the user to visually explore the data as part of the analytical process whereas our analytical brushing method allows for point-and-click analysis and then utilizes visual exploration for refinement. In this way, the framework completes the visual analytics pipeline of "analyze first", "show the important", "zoom and filter", "analyze further", "details on demand" [3]. Furthermore, the combination of similarity metrics with a linked visualization tool provides users with a means of directly assessing the similarity results. By enabling users to design their own temporal similarity

functions, this framework allows them to utilize their domain knowledge and inject information into the system that may otherwise be difficult to capture. As was seen in the case studies, similarity does not imply equality, nor does it imply future trends. In some temporal plots, the graphs may be relatively similar; however, trajectories may be trending slightly downwards towards the end of the time series of some similar regions and trending slightly upwards in other similar regions. Without a linked visualization, the only resulting metric for analysis would be the similarity score. By directly linking both the analytical component and visual component together, this framework is able to provide users with a combined tool that is stronger than its individual components.

Moreover, this thesis presents an interactive geovisual analytics framework which allows users to explore the impact of geographical variations across locations and scales for multivariate data clustering. The space has been categorized into four aspects: discrete spatial extent, discrete geographical features, continuous spatial extent, and continuous geographical resolution in order to characterize the impact of spatial dependence and heterogeneity. A variety of visualization and interaction techniques (e.g., PCA scatterplot, PCP area profiler, Rose plot) have been implemented to facilitate clustering exploration over geographical variations with statistical measures (e.g., Silhouette coefficient) to evaluate cluster quality. This framework also provides methods for comparing within (k-means vs. k-means) and between (hierarchical vs. k-means) cluster results, and demonstrate potential ways of interacting with data to explore cluster results.

Last but not least, another major contribution of this thesis is an additional metric that can be assessed during map design. A critical step in designing choropleth maps is the choice of classification method. How a map is classified directly impacts the resultant visual output and can lead to misinformation about the underlying data. In

order to assess the visual impact of such choices, this thesis has developed a metric for quantifying the visual impact of adjusting classification boundaries in a choropleth map. Based on this metric a scheme is presented for maximizing or minimizing the amount of visual clustering present in the map and demonstrated the results using several datasets. What is critical to note is that the goal of choosing classification boundaries is to achieve a reasonable split in the data, and this is often left up to the designer. By providing designers with new ways to assess the visual impact of small classification changes, the designer can further refine and assess their map message.

There are many extensions for this thesis that are worth further development. First of all, the current framework has many Exploration Data Analysis features to enable analysts to quickly identify the points of interest. However, there are a variety of potential extensions. For example, one can envision a palette of modifiable analytical brushes which experts could use to paint their data, extract interesting features, and form hypotheses. The framework could also be extended to have future brushes that could highlight spatiotemporal cluster stability, or other advanced metrics. Moreover, while this framework enables the exploration and comparison of clustering methods over different scales, there is still a need to enable quick identification of similar and dissimilar regions. Currently, the comparative analysis between clusters is done in a purely visual manner, and while humans are capable of identifying patterns, the integration of further analytical methods to help highlight and identify statistically significant similarities and differences between clusters is critical. Furthermore, this exploration focused primarily on the spatial extent of the data; however, extensions to the spatiotemporal domain are critical in analyzing how underlying physical properties may develop in the data. It may also be possible to automatically explore the impact of scale simply by defining levels of aggregation and present a summary comparison to end users to suggest appropriate scales of anal-

ysis for the data. Future work should explore a combination of automation with human-in-the-loop exploration and recommendations.

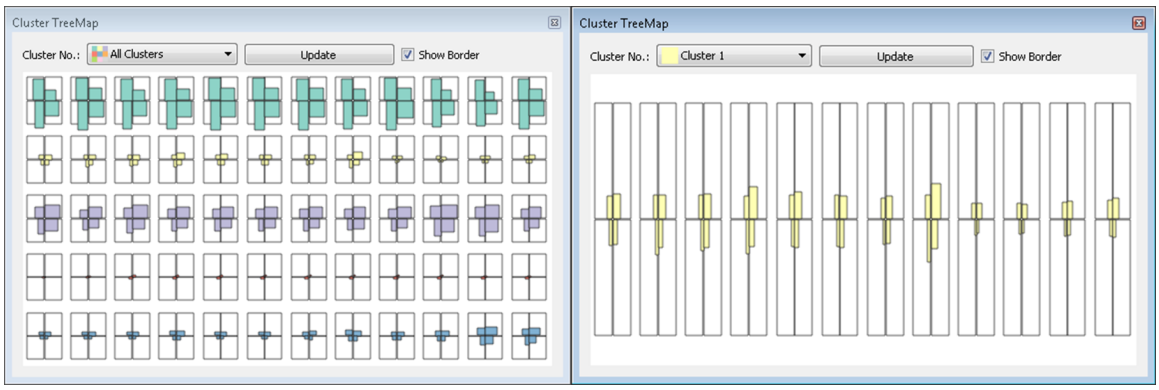


Figure 8.1: An example of a temporal tree map for 5 clusters over 12 years. The left part is an overview of all 5 clusters and the right part is a selected view for the yellow cluster.

To better analyze the spatiotemporal clusters, a cluster tree map that visualizes the temporal change of clusters in space could be designed. After the temporal coherent clustering process, the tree map may divide the geographic space into four quadrants. For each quadrant, the ratio of each cluster size over the number of spatial units in that quadrant can be calculated and represented by the area of its corresponding tree map branch. The spatial distribution for each cluster can be aggregated and represented as the aspect ratio of its corresponding tree map branch (Figure 8.1). For instance, if the members of a certain cluster in the third quadrant are distributed more in the south direction than the west direction, that tree map branch should also be extending in the south direction. In this way, users can easily identify the consistency of clusters over time.

As for quantifying the visual impact of classification boundaries, future research could explore the sensitivity of such an application across various clustering schemes. Research into what pattern changes result in an increased perception of visual clus-

tering should also be undertaken. Currently, we rely solely on the fact that colors are changing and regions are becoming larger. Past research [203] has shown that the larger a patch of color becomes in a choropleth map, the more likely that it will be identified as a cluster. However, there may be particular patches that could be changed that may have a greater impact on the perception. Of course, the size of the spatial unit matters a great deal, but what if changing the classification of a spatial unit fills in a donut hole? Is this perceived as resulting in more spatial clustering than if we change a spatial unit's classification such that it just adds to the edge of the donut? What if a spatial unit acts as a bridge? For example, if there are two spatial groupings with the same label separated by a narrow band of other labels, how is the spatial clustering perceived if one unit is changed to create a bridge? Understanding the impact of these patterns would allow us to computationally identify them and use these types of patterns to create a more perceptually rigorous VIOC metric. Future work can focus on the use of such metrics for highlighting uncertainty within the map, as well as exploring boundary elements with respect to statistical measures of spatial clustering. Specifically, if a region is found to be a statistically significantly spatially cluster, should boundary elements contiguous to this region be adjusted to highlight the significance?

REFERENCES

[1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, Tech. Rep., 2011.

[2] M. Batty and J. Cheshire, "Cities as flows, cities of flows," *Environment and Planning B: Planning and Design*, vol. 38, no. 2, pp. 195–196, 2011.

[3] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis," in *Tenth International Conference on Information Visualization*. IEEE, 2006, pp. 9–16.

[4] K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep., 2005.

[5] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[6] C. A. Brewer and L. Pickle, "Evaluation of methods for classifying epidemiological data on choropleth maps in series," *Annals of the Association of American Geographers*, vol. 92, no. 4, pp. 662–681, 2002.

[7] J. Zhang and H. Shi, "Geo-visualization and clustering to support epidemiology surveillance exploration," in *International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2010, pp. 381–386.

[8] R. Maciejewski, R. Hafen, S. Rudolph, S. Larew, M. Mitchell, W. Cleveland, and D. Ebert, "Forecasting hotspots - a predictive analytics approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 4, pp. 440–453, 2011.

[9] C. Bryan, X. Wu, S. Mniszewski, and K.-L. Ma, "Integrating predictive analytics into a spatiotemporal epidemic simulation," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2015, pp. 17–24.

[10] C. C. Calhoun, C. E. Stobbart, D. M. Thomas, J. A. Villarrubia, D. E. Brown, and J. H. Conklin, "Improving crime data sharing and analysis tools for a web-based crime analysis toolkit: WebCAT 2.2," in *Systems and Information Engineering Design Symposium*. IEEE, 2008, pp. 40–45.

[11] A. Malik, R. Maciejewski, T. F. Collins, and D. S. Ebert, "Visual analytics law enforcement toolkit," in *Proceedings of the IEEE Conference on Technologies for Homeland Security*, 2010, pp. 222–228.

[12] A. M. Razip, A. Malik, S. Afzal, M. Potrawski, R. Maciejewski, Y. Jang, N. Elmqvist, and D. S. Ebert, "A mobile visual analytics approach for law enforcement situation awareness," in *IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2014, pp. 169–176.

[13] W. W. Hargrove and F. M. Hoffman, "Potential of multivariate quantitative methods for delineation and visualization of ecoregions," *Environmental Management*, vol. 34, pp. S39–S60, 2005.

[14] F. M. Hoffman, W. W. Hargrove, R. T. Mills, S. Mahajan, D. J. Erickson, and R. J. Oglesby, "Multivariate spatio-temporal clustering (MSTC) as a data mining tool for environmental applications," in *Proceedings of the international congress on environmental modelling and software society fourth biennial meeting:*, 2008, pp. 1774–1781.

[15] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, D. S. Ebert, and W. Huang, "A correlative analysis process in a visual analytics environment," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2012, pp. 33–42.

[16] R. T. Mills, F. M. Hoffman, J. Kumar, and W. W. Hargrove, "Cluster analysis-based approaches for geospatiotemporal data mining of massive data sets for identification of forest threats," *Procedia Computer Science*, vol. 4, pp. 1612–1621, 2011.

[17] D. I. Ashby and P. A. Longley, "Geocomputation, geodemographics and resource allocation for local policing," *Transactions in GIS*, vol. 9, no. 1, pp. 53–72, 2005.

[18] A. Savikhin, R. Maciejewski, and D. S. Ebert, "Applied visual analytics for economic decision-making," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2008, pp. 107–114.

[19] W. Luo, A. M. MacEachren, P. Yin, and F. Hardisty, "Spatial-social network visualization for exploratory data analysis," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 2011, pp. 65–68.

[20] M. Kraak and F. Ormeling, *Cartography: Visualization of Geospatial Data*. Pearson Education, 2003. [Online]. Available: https://books.google.com/books?id=7rsC3FeJhkkC

[21] L. Anselin, I. Syabri, and Y. Kho, "GeoDa: An introduction to spatial data analysis," *Geographical Analysis*, vol. 38, no. 1, pp. 5–22, 2006.

[22] N. Cao, D. Gotz, J. Sun, and H. Qu, "Dicon: Interactive visual analysis of multidimensional clusters," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2581–2590, 2011.

[23] D. Guo, J. Chen, A. M. MacEachren, and K. Liao, "A visualization system for space-time and multivariate patterns (VIS-STAMP)," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1461–1474, Nov. 2006.

[24] P. Gatalsky, N. Andrienko, and G. Andrienko, "Interactive analysis of event data using space-time cube," in *Eighth International Conference on Information Visualisation*. IEEE, 2004, pp. 145–152.

[25] S. Kim, R. Maciejewski, A. Malik, Y. Jang, D. S. Ebert, and T. Isenberg, "Bristle maps: A multivariate abstraction technique for geovisualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 9, pp. 1438–1454, 2013.

[26] S. Dbel, M. Rhlig, H. Schumann, and M. Trapp, "2D and 3D presentation of spatial data: A systematic review," in *2014 IEEE VIS International Workshop on 3DVis*. IEEE, 2014, pp. 11–18.

[27] X. Liu, Y. Hu, S. North, and H.-W. Shen, "Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling," *Computer Graphics Forum*, 2015.

[28] K. Goldsberry and S. Battersby, "Issues of change detection in animated choropleth maps," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 44, no. 3, pp. 201–215, 2009.

[29] L. Anselin, "Local indicators of spatial association–LISA," *Geographical analysis*, vol. 27, no. 2, pp. 93–115, 1995.

[30] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, pp. 234–240, 1970.

[31] M. Monmonier, *How to lie with maps*. University of Chicago Press, 2014.

[32] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data." in *Pacific symposium on biocomputing*, vol. 7, no. 6. World Scientific, 2002, pp. 6–17.

[33] Y. Hu and R. J. Hathaway, "An algorithm for clustering tendency assessment," *WSEAS Trans. Math*, vol. 7, no. 7, pp. 441–450, 2008.

[34] U. Von Luxburg, *Clustering stability: An overview*. Now Publishers Inc, 2010.

[35] C. Hennig, "Cluster-wise assessment of cluster stability," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 258–271, 2007.

[36] S. Imfeld, "Time, points and space-towards a better analysis of wildlife data in gis," Ph.D. dissertation, Geographisches Institut der Universität Zürich, 2000.

[37] M. Erwig, "Toward spatio-temporal patterns," in *Spatio-Temporal Databases*. Springer, 2004, pp. 29–53.

[38] L. Anselin and S. Bao, "Exploratory spatial data analysis linking SpaceStat and ArcView," *Recent Developments in Spatial Analysis—Spatial Statistics, Behavioural Modelling and Computational Intelligence*, pp. 35–59, 1997.

[39] L. Anselin, R. Dodson, and S. Hudak, "Linking GIS and spatial data analysis in practice," *Geographical Systems*, vol. 1, pp. 3–23, 1993.

[40] A. Unwin, G. Wills, and J. Haslett, "REGARD - graphical analysis of regional data," in *Proceedings of the Section on Statistical Graphics*. Alexandria, VA: American Statistical Association, 1990, pp. 36–41.

[41] G. Wills, A. R. Unwin, and J. Haslett, "Spatial interactive graphics applied to Irish socioeconomic data," in *Proceedings of the ASA Statistical Graphics Section*. Atlanta, Georgia: American Statistical Association, 1991, pp. 37–41.

[42] J. Haslett, G. Wills, and A. Unwin, "SPIDER–interactive statistical tool for the analysis of spatially distributed data," *International journal of geographical information systems*, vol. 4, no. 3, pp. 285–296, 1990.

[43] D. F. Swayne, D. Cook, and A. Buja, "Xgobi: Interactive Dynamic Graphics In The X Window System With A Link To S," 1991.

[44] J. Symanzik, T. Kötter, S. Schmelzer, S. Klinke, D. Cook, and D. F. Swayne, "Spatial Data Analysis in the Dynamically Linked ArcView/XGobi/XploRe Environment," *Computing Science and Statistics*, vol. 29, pp. 561–569, 1997.

[45] J. Symanzik, D. Cook, N. Lewin-Koh, J. J. Majure, and I. Megretskaia, "Linking ArcView^{TM} and XGobi: Insight behind the Front End," *Journal of Computational and Graphical Statistics*, vol. 9, no. 3, pp. 470–490, 2000.

[46] M. O. Ward, "XmdvTool: integrating multiple methods for visualizing multivariate data," in *Proceedings of the conference on Visualization '94*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1994, pp. 326–333.

[47] E. A. Rundensteiner, M. O. Ward, Z. Xie, Q. Cui, C. V. Wad, D. Yang, and S. Huang, "Xmdvtool^{Q}:: quality-aware interactive data exploration," in *SIGMOD Conference*, 2007, pp. 1109–1112.

[48] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics-Theory and methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[49] M. Kulldorff, K. Rand, G. Gherman, G. Williams, and D. DeFrancesco, "SaTScan v 2.1: Software for the spatial and space-time scan statistics," *Bethesda, MD: National Cancer Institute*, 1998.

[50] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari, "A space–time permutation scan statistic for disease outbreak detection," *PLoS Medicine*, vol. 2, no. 3, p. e59, 2005.

[51] S. Wise, R. Haining, and J. Ma, "Regionalization tools for exploratory spatial analysis of health data," *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling, and computational intelligence*, pp. 83–100, 1997.

[52] R. Haining, S. Wise, and J. Ma, "Designing and implementing software for spatial statistical analysis in a GIS environment," *Journal of Geographical Systems*, vol. 2, pp. 257–286, 2000.

[53] S. Wise, R. Haining, and J. Ma, "Providing spatial statistical data analysis functionality for the GIS user: the SAGE project," *International Journal of Geographical Information Science*, vol. 15, no. 3, pp. 239–254, 2001.

[54] J. Dykes, "Pushing maps past their established limits: a unified approach to cartographic visualization," *Innovations in GIS*, vol. 3, pp. 177–87, 1995.

[55] ——, "Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using TcI/Tk and cdv," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 47, no. 3, pp. 485–497, 1998.

[56] N. Levine, "CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations," in *The IV International Conference on GeoComputation*, Fredericksburg, July 1999.

[57] ——, "Crime mapping and the CrimeStat program," *Geographical Analysis*, vol. 38, no. 1, pp. 41–55, 2006.

[58] ——, *CrimeStat: A spatial statistics program for the analysis of crime incident locations(V 3.3)*, Electronically, National Institute of Justice, Washington, DC, Jul. 2010.

[59] M. Gahegan, M. Takatsuka, M. Wheeler, and F. Hardisty, "Abstract GeoVISTA studio: A geocomputational workbench," in *The Proceedings of the 5th International Conference on GeoComputation*, August 2000.

[60] M. Takatsuka and M. Gahegan, "GeoVISTA studio: A codeless visual programming environment for geoscientific data analysis and visualization," *Computational Geoscience*, vol. 28, pp. 1131–1144, 2002.

[61] M. Takasuka and M. Gahegan, "Sharing exploratory geospatial analysis and decision making using GeoVISTA studio: From a desktop to the web," *Journal of Geographic Information and Decision Analysis*, vol. 5, pp. 129–139, 2001.

[62] R. Bivand and A. Gebhardt, "Implementing functions for spatial statistical analysis using the R language," *Journal of Geographical Systems*, vol. 2, pp. 307–317, 2000.

[63] R. Bivand, "Spatial econometrics functions in R: Classes and methods," *Journal of Geographical Systems*, vol. 4, pp. 405–421, 2002.

[64] L. Anselin, *An Introduction to Spatial Regression Analysis in R*, University of Illinois, Urbana-Champaign, 2003.

[65] ——, *An introduction to EDA with GeoDa*, Spatial Analysis Laboratory, Department of Agricultural and Consumer Economics, UIUC, June 2003.

[66] V. Gomez-Rubio, J. Ferrandiz, and A. Lopez, "Detecting clusters of diseases with R," *Journal of geographical systems*, vol. 7, pp. 189–206, 2003.

[67] C. Weaver, "Building Highly-Coordinated Visualizations in Improvise," in *Proceedings of the IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 159–166.

[68] C. E. Weaver, "Improvise: A user interface for interactive construction of highly-coordinated visualizations," Ph.D. dissertation, University of Wisconsin at Madison, Madison, WI, USA, 2006.

[69] J. LeSage and R. Pace, "Arc_Mat, a toolbox for using ArcView shape files for spatial econometrics and statistics," in *Geographic Information Science*, ser. Lecture Notes in Computer Science, M. Egenhofer, C. Freksa, and H. Miller, Eds. Springer Berlin Heidelberg, 2004, vol. 3234, pp. 179–190. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30231-5_12

[70] X. Liu and J. LeSage, "Arc_Mat: A matlab-based spatial data analysis toolbox," *Journal of geographical systems*, vol. 12, no. 1, pp. 69–87, 2010.

[71] S. J. Rey and M. V. Janikas, "STARS: Space–time analysis of regional systems," *Geographical analysis*, vol. 38, no. 1, pp. 67–86, 2006.

[72] M. V. Janikas and S. J. Rey, "Spatial clustering, inequality and income convergence," *Région et Développement*, vol. 21, pp. 45–64, 2005.

[73] C. Robertson, T. Nelson, B. Boots, and M. Wulder, "STAMP: spatial–temporal analysis of moving polygons," *Journal of Geographical Systems*, vol. 9, no. 3, pp. 207–227, September 2007.

[74] J. McIntosh and M. Yuan, "A framework to enhance semantic flexibility for analysis of distributed phenomena," *International Journal of Geographical Information Science*, vol. 19, no. 10, pp. 999–1018, 2005.

[75] Y. Sadahiro and M. Umemura, "A computational approach for the analysis of changes in polygon distributions," *Journal of Geographical Systems*, vol. 3, no. 2, pp. 137–154, 2001.

[76] E. A. Mack, Y. Zhang, S. Rey, and R. Maciejewski, "Spatio-temporal analysis of industrial composition with IVIID: An interactive visual analytics interface for industrial diversity," *Journal of Geographical Systems*, vol. 16, no. 2, pp. 183–209, 2014.

[77] J. Symanzik, J. Majure, and D. Cook, "Dynamic graphics in a GIS: A bidirectional link between ArcView 2.0 and XGobi," *Computing Science and Statistics*, vol. 27, pp. 299–303, 1996.

[78] B. Everett, *Cluster Analysis*. London: Heinemann Educational Books Ltd, 1974.

[79] A. McBratney and J. Gruijter, "A continuum approach to soil classification by modified fuzzy k-means with extragrades," *Journal of Soil Science*, vol. 43, no. 1, pp. 159–175, 2006.

[80] P. Moran, "The interpretation of statistical maps," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10, no. 2, pp. 243–251, 1948.

[81] R. Geary, "The contiguity ratio and statistical mapping," *The incorporated statistician*, vol. 5, no. 3, pp. 115–146, 1954.

[82] A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical analysis*, vol. 24, no. 3, pp. 189–206, 1992.

[83] L. Anselin, "Space-time mapping," July 2012, Spatial Autocorrelation, Chicago, IL .

[84] R. Block and C. Block, "Space, place and crime: Hot spot areas and hot places of liquor-related crime," *Crime and place*, vol. 4, no. 2, pp. 145–184, 1995.

[85] E. Knox and M. Bartlett, "The detection of space-time interactions," *Applied Statistics*, pp. 25–30, 1964.

[86] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[87] S. Rey, L. Anselin, X. Li, and J. Koschinsky, "Guide to using the crime analytics for spacetime," Tech. Rep., 2013.

[88] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 3–10.

[89] G. L. Andrienko, N. V. Andrienko, I. Kopanakis, G. Marketos, and Y. Theodoridis, "Visually exploring movement data via similarity-based analysis," *Journal of Intelligent Inforrmation Systems*, vol. 38, pp. 343–391, 2012.

[90] C. Tominski, H. Schumann, G. Andrienko, and N. Andrienko, "Stacking-based visualization of trajectory attribute data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2565–2574, 2012.

[91] N. Pelekis, G. Andrienko, N. Andrienko, I. Kopanakis, G. Marketos, and Y. Theodoridis, "Visually exploring movement data via similarity-based analysis," *Journal of Intelligent Information Systems*, vol. 38, no. 2, pp. 343–391, 2012.

[92] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2149–2158, 2013.

[93] C. Zhong, T. Wang, W. Zeng, and S. M. Arisona, "Spatiotemporal visualisation: A survey and outlook," in *Digital Urban Modeling and Simulation*. Springer, 2012, pp. 299–317.

[94] N. Andrienko and G. Andrienko, "Interactive visual tools to explore spatio-temporal variation," in *Proceedings of the working conference on Advanced visual interfaces*, 2004, pp. 417–420.

[95] J. Dykes, A. MacEachren, and M. Kraak, "Impact of data and task characteristics on design of spatio-temporal data visualization tools," *Exploring geovisualization*, p. 201, 2005.

[96] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003.

[97] E. L. Koua, A. MacEachren, and M.-J. Kraak, "Evaluating the usability of visualization methods in an exploratory geovisualization environment," *International journal of geographical information science*, vol. 20, no. 4, pp. 425–448, 2006.

[98] C. Daassi, L. Nigay, and M.-C. Fauvet, "A taxonomy of temporal data visualization techniques," *Information-Interaction-Intelligence*, vol. 5, no. 2, pp. 41–63, 2005.

[99] E. A. Mack and R. Maciejewski, "A profile of visual analytical toolkits for understanding the spatio-temporal evolution of broadband provision," *Telecommunications Policy*, vol. 39, no. 3, pp. 320–332, 2015.

[100] B. Tversky, J. B. Morrison, and M. Bertrancourt, "Animation: Can it facilitate?" *International Journal of Human-Computer Studies*, vol. 57, pp. 247–262, 2002.

[101] L. Chittaro and C. Combi, "Visualizing queries on databases of temporal histories: New metaphors and their evaluation," *Data & Knowledge Engineering*, vol. 44, no. 2, pp. 239–264, 2003.

[102] M. C. Hao, U. Dayal, D. A. Keim, and T. Schreck, "Importance-driven visualization layouts for large time series data," in *Proceedings of the IEEE Symposium on Information Visualization*, 2005, pp. 203–210.

[103] J. Lin, E. Keogh, and S. Lonardi, "Visualizing and discovering non-trivial patterns in large time series databases," *Information Visualization*, vol. 4, no. 2, pp. 61–82, 2005.

[104] R. Bade, S. Schlechtweg, and S. Miksch, "Connecting time-oriented data and information to a coherent interactive visualization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 105–112.

[105] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualization*, vol. 3, no. 1, pp. 1–18, 2004.

[106] K. Wongsuphasawat, C. Plaisant, M. Taieb-Maimon, and B. Shneiderman, "Querying event sequences by exact match or similarity search: Design and empirical evaluation," *Interacting with Computers*, vol. 24, no. 2, pp. 55–68, 2012.

[107] A. Alencar, F. Paulovich, R. Minghim, M. Filho, and M. Oliveira, "Similarity-based visualization of time series collections: An application to analysis of streamflows," in *Proceedings of the 12th International Conference Information Visualisation*, 2008, pp. 280 –286.

[108] O. Hoeber, G. Wilson, S. Harding, R. Enguehard, and R. Devillers, "Exploring geo-temporal differences using GTdiff," in *Pacific Visualization Symposium*. IEEE, 2011, pp. 139–146.

[109] G. Andrienko and N. Andrienko, "Visual exploration of the spatial distribution of temporal behaviors," in *Ninth International Conference on Information Visualisation*. IEEE, 2005, pp. 799–806.

[110] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[111] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment, 2004, pp. 792–803.

[112] L. Axis Maps, "Indiemapper," 2010.

[113] P. Longley, *Geographic information systems and science*. John Wiley & Sons, 2005.

[114] G. F. Jenks, "The data model concept in statistical mapping," *International yearbook of cartography*, vol. 7, no. 1, pp. 186–190, 1967.

[115] M. W. Scripter, "Nested-means map classes for statistical maps," *Annals of the Association of American Geographers*, vol. 60, no. 2, pp. 385–392, 1970.

[116] E. Cromley and R. Cromley, "An analysis of alternative classification schemes for medical atlas mapping," *European Journal of Cancer*, vol. 32, no. 9, pp. 1551–1559, 1996.

[117] M. P. Armstrong, N. Xiao, and D. A. Bennett, "Using genetic algorithms to create multicriteria class intervals for choropleth maps," *Annals of the Association of American Geographers*, vol. 93, no. 3, pp. 595–623, 2003.

[118] I. S. Evans, "The selection of class intervals," *Transactions of the Institute of British Geographers*, pp. 98–124, 1977.

[119] M. Sun, D. Wong, and B. Kronenfeld, "A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping," *Cartography and Geographic Information Science*, pp. 1–13, 2016.

[120] D. Vickers and P. Rees, "Creating the UK National Statistics 2001 output area classification," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 2, pp. 379–403, 2007.

[121] L. Scrucca, "Clustering multivariate spatial data based on local measures of spatial autocorrelation," Universit di Perugia, Dipartimento Economia, Quaderni del Dipartimento di Economia, Finanza e Statistica 20/2005, Oct. 2005. [Online]. Available: https://ideas.repec.org/p/pia/wpaper/20-2005.html

[122] G. Andrienko, N. Andrienko, and A. Savinov, "Choropleth maps: classification revisited," in *Proceedings ICA*. Citeseer, 2001, pp. 1109–1219.

[123] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, and D. Keim, "Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns," in *Computer Graphics Forum*, vol. 29, no. 3. Wiley Online Library, 2010, pp. 913–922.

[124] M. Streit, H.-J. Schulz, A. Lex, D. Schmalstieg, and H. Schumann, "Model-driven design for the visual analysis of heterogeneous data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 6, pp. 998–1010, 2012.

[125] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[126] R. Harris, P. Sleight, and R. Webber, *Geodemographics, GIS and neighbourhood targeting*. John Wiley and Sons, 2005, vol. 7.

[127] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[128] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[129] F. K. Kuiper and L. Fisher, "391: A monte carlo comparison of six clustering procedures," *Biometrics*, pp. 777–783, 1975.

[130] L. Ferreira and D. B. Hitchcock, "A comparison of hierarchical methods for clustering functional data," *Communications in Statistics-Simulation and Computation*, vol. 38, no. 9, pp. 1925–1949, 2009.

[131] S. A. Fattah, C.-C. Lin, and S.-Y. Kung, "A mutual information based approach for evaluating the quality of clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 601–604.

[132] Y. Jung, H. Park, D.-Z. Du, and B. L. Drake, "A decision criterion for the optimal number of clusters in hierarchical clustering," *Journal of Global Optimization*, vol. 25, no. 1, pp. 91–111, 2003.

[133] M. Meilă, "Comparing clusterings: An axiomatic view," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 577–584.

[134] F. M. Hoffman, W. W. Hargrove, and A. D. Del Genio, "Multivariate spatiotemporal clustering of time-series data: An approach for diagnosing cloud properties and understanding ARM site representativeness," in *Thirteenth ARM Science Team Meeting Proc., Broomfield, Colorado*, 2003.

[135] J. Zhou, S. Konecni, and G. Grinstein, "Visually comparing multiple partitions of data with applications to clustering," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2009, pp. 72 430J–72 430J.

[136] Y. Hu, S. G. Kobourov, and S. Veeramoni, "Embedding, clustering and coloring for dynamic maps," in *IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2012, pp. 33–40.

[137] L. Anselin, *Spatial econometrics: Methods and models*. Springer Science & Business Media, 1988, vol. 4.

[138] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically weighted regression: A method for exploring spatial nonstationarity," *Geographical analysis*, vol. 28, no. 4, pp. 281–298, 1996.

[139] L. Anselin, *The Moran scatterplot as an ESDA tool to assess local instability in spatial association*. Regional Research Institute, West Virginia University Morgantown, WV, 1993.

[140] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 47, no. 3, pp. 431–443, 1998.

[141] J. K. Ord and A. Getis, "Local spatial autocorrelation statistics: distributional issues and an application," *Geographical analysis*, vol. 27, no. 4, pp. 286–306, 1995.

[142] C. Brunsdon, A. Fotheringham, and M. Charlton, "Geographically weighted summary statistics - a framework for localised exploratory data analysis," *Computers, Environment and Urban Systems*, vol. 26, no. 6, pp. 501–524, 2002.

[143] P. Harris, C. Brunsdon, and M. Charlton, "Geographically weighted principal components analysis," *International Journal of Geographical Information Science*, vol. 25, no. 10, pp. 1717–1736, 2011.

[144] J. Dykes and C. Brunsdon, "Geographically weighted visualization: Interactive graphics for scale-varying exploratory analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1161–1168, 2007.

[145] C. Turkay, A. Slingsby, H. Hauser, J. Wood, and J. Dykes, "Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2033–2042, 2014.

[146] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay, "Visualizing multiple variables across scale and geography," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 599–608, 2016.

[147] E. Mack, T. H. Grubesic, and E. Kessler, "Indices of industrial diversity and regional economic composition," *Growth and Change*, vol. 38, no. 3, pp. 474–509, 2007.

[148] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: A brief survey," *A review of machine learning techniques for processing multimedia content*, p. 11, 2004.

[149] D. Cohn, R. Caruana, and A. McCallum, "Semi-supervised clustering with user feedback," *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, vol. 4, no. 1, pp. 17–32, 2003.

[150] A. K. Jain, P. K. Mallapragada, and M. Law, "Bayesian feedback in data clustering," in *18th International Conference on Pattern Recognition*, vol. 3. IEEE, 2006, pp. 374–378.

[151] Y. Huang and T. M. Mitchell, "Text clustering with extended user feedback," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 413–420.

[152] M.-F. Balcan and A. Blum, "Clustering with interactive feedback," in *Algorithmic Learning Theory*. Springer, 2008, pp. 316–328.

[153] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Weakly supervised nonnegative matrix factorization for user-driven clustering," *Data Mining and Knowledge Discovery*, pp. 1–24, 2014.

[154] K. Chen and L. Liu, "VISTA: Validating and refining clusters via visualization," *Information Visualization*, vol. 3, no. 4, pp. 257–270, 2004.

[155] ——, "iVIBRATE: Interactive visualization-based framework for clustering large datasets," *ACM Transactions on Information Systems (TOIS)*, vol. 24, no. 2, pp. 245–294, 2006.

[156] L. House, S. Leman, and C. Han, "Bayesian visual analytics (BaVA)," *FODAVA Technical Report*, 2010.

[157] S. C. Leman, L. House, D. Maiti, A. Endert, and C. North, "Visual to parametric interaction (V2PI)," *PloS one*, vol. 8, no. 3, p. e50474, 2013.

[158] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North, "Observation-level interaction with statistical models for visual analytics," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2011, pp. 121–130.

[159] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning distance functions interactively," in *IEEE Conference on Visual Analytics Science and Technology*. IEEE, 2012, pp. 83–92.

[160] C. G. Healey and J. T. Enns, "Attention and visual memory in visualization and computer graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 7, pp. 1170–1188, 2012.

[161] R. A. Becker and W. S. Cleveland, "Brushing scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, 1987.

[162] G. Dang, C. North, and B. Shneiderman, "Dynamic queries and brushing on choropleth maps," in *Proceedings of the Fifth International Conference on Information Visualisation*, 2001, pp. 757–764.

[163] J. Talbot, S. Lin, and P. Hanrahan, "An extension of Wilkinson's algorithm for positioning tick labels on axes," *IEEE Transactions on Visualization and Computer Graphics*, 2010.

[164] C. A. Brewer, *Designing better Maps: A Guide for GIS users*. ESRI Press, 2005.

[165] A. C. Robinson, "Highlighting in geovisualization," *Cartography and Geographic Information Science*, vol. 38, no. 4, pp. 373–383, 2011.

[166] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, "Research directions in data wrangling: Visualizations and transformations for usable and credible data," *Information Visualization*, vol. 10, no. 4, pp. 271–288, 2011.

[167] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 3363–3372.

[168] J. W. Tukey, "Exploratory data analysis," *Reading, Ma*, vol. 231, p. 32, 1977.

[169] L. Anselin, "Interactive techniques and exploratory spatial data analysis," *Geographical Information Systems: Principles, Techniques, Management and Applications, eds., P. Longley, M. Goodchild, D. Maguire, and D. Rhind. Cambridge: Geoinformation Int*, 1999.

[170] D. J. Kasik, D. Ebert, G. Lebanon, H. Park, and W. M. Pottenger, "Data transformations and representations for computation and visualization," *Information Visualization*, vol. 8, no. 4, pp. 275–285, 2009.

[171] V. R. Patel and R. G. Mehta, "Impact of outlier removal and normalization approach in modified k-means clustering algorithm," *International Journal of Computer Science Issues*, vol. 8, no. 5, 2011.

[172] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[173] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[174] M. Harrower, "The cognitive limits of animated maps," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 42, no. 4, pp. 349–357, 2009.

[175] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.

[176] J. Bernard, J. Brase, D. Fellner, O. Koepler, J. Kohlhammer, T. Ruppert, T. Schreck, and I. Sens, "A visual digital library approach for time-oriented scientific primary data," *International Journal on Digital Libraries*, vol. 11, no. 2, pp. 111–123, 2010.

[177] M. C. Hao, U. Dayal, D. Keim, D. Morent, J. Schneidewind *et al.*, "Intelligent visual analytics queries," in *IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2007, pp. 91–98.

[178] J. Choo, H. Lee, J. Kihm, and H. Park, "iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2010.

[179] M. Hossain, P. Ojili, C. M. Grimm, R. Muller, L. T. Watson, and N. Ramakrishnan, "Scatter/gather clustering: Flexibly incorporating user feedback to steer clustering results," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2829 – 2838, 2012.

[180] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*. Springer, 2006, pp. 25–71.

[181] M. Adnan, A. Singleton, and P. Longley, "Spatially weighted geodemographics," *GIS Research UK 21st Annual Conference*, 2013.

[182] A. Slingsby, J. Dykes, and J. Wood, "Rectangular hierarchical cartograms for socio-economic data," *Journal of Maps*, vol. 6, no. 1, pp. 330–345, 2010.

[183] ——, "Exploring uncertainty in geodemographics with interactive graphics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2545–2554, 2011.

[184] G. Grekousis and H. Thomas, "Comparison of two fuzzy algorithms in geodemographic segmentation analysis: The fuzzy c-means and gustafson–kessel methods," *Applied Geography*, vol. 34, pp. 125–136, 2012.

[185] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.

[186] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[187] C. Gini, *Variabilità e mutabilità*, 1912.

[188] B. Shneiderman, "The future of interactive systems and the emergence of direct manipulation," *Behaviour & Information Technology*, vol. 1, no. 3, pp. 237–256, 1982.

[189] ——, "Direct manipulation: A step beyond programming languages," in *ACM SIGSOC Bulletin*, vol. 13, no. 2-3. ACM, 1981, p. 143.

[190] S. Openshaw, *The modifiable areal unit problem*. Norwick [Norfolk]: Geo Books, 1983.

[191] J. E. Burt, G. M. Barber, and D. L. Rigby, *Elementary statistics for geographers*. Guilford Press, 2009.

[192] T. Schreck, I. Omer, P. Bak, and Y. Lerman, *Geographic Information Science at the Heart of Europe*. Cham: Springer International Publishing, 2013, ch. A Visual Analytics Approach for Assessing Pedestrian Friendliness of Urban Environments, pp. 353–368.

[193] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial & Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[194] G. J. Torres, R. B. Basnet, A. H. Sung, S. Mukkamala, and B. M. Ribeiro, "A similarity measure for clustering and its applications," *Int. J. of Elec. Comput. & Syst. Eng*, vol. 3, pp. 164–170, 2009.

[195] A. Klippel, F. Hardisty, and R. Li, "Interpreting spatial patterns: An inquiry into formal and cognitive aspects of tobler's first law of geography," *Annals of the Association of American Geographers*, vol. 101, no. 5, pp. 1011–1031, 2011.

[196] M. H. Kalos and P. A. Whitlock, *Monte carlo methods*. John Wiley & Sons, 2008.

[197] A. D. Cliff and J. K. Ord, *Spatial autocorrelation*. Pion London, 1973, vol. 5.

[198] P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, pp. 17–23, 1950.

[199] A. D. Cliff and K. Ord, "Spatial autocorrelation: A review of existing and new measures with applications," *Economic Geography*, vol. 46, pp. 269–292, 1970.

[200] L. J. Hubert, R. G. Golledge, and C. M. Costanzo, "Generalized procedures for evaluating spatial autocorrelation," *Geographical Analysis*, vol. 13, no. 3, pp. 224–233, 1981.

[201] E. M. Macambira and C. C. de Souza, "The edge-weighted clique problem: Valid inequalities, facets and polyhedral computations," *European Journal of Operational Research*, vol. 123, no. 2, pp. 346–371, 2000.

[202] U. Faigle, W. Kern, and G. Still, *Algorithmic principles of mathematical programming*. Springer Science & Business Media, 2013, vol. 24.

[203] M. Haklay, *Interacting with geospatial technologies*. Wiley Online Library, 2010.

[204] S. Moon, E.-K. Kim, and C.-S. Hwang, "Effects of spatial distribution on change detection in animated choropleth maps," *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, vol. 32, no. 6, pp. 571–580, 2014.

[205] A. F. Tollefsen, H. Strand, and H. Buhaug, "PRIO-GRID: A unified spatial data structure," *Journal of Peace Research*, vol. 49, no. 2, pp. 363–374, 2012.

[206] "North american industry classification system," 2010, [Accessed: 2015-03-29]. [Online]. Available: http://www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2012

[207] S. Matthews, "Four states hit hardest by housing now lead U.S. jobs recovery," 2012. [Online]. Available: http://www.bloomberg.com/news/articles/2012-03-09/states-hardest-hit-by-real-estate-collapse-lead-u-s-labor-market-recovery

[208] W. Cox, "How Texas avoided the great recession," 2010, [Online; posted 20-Jul-2010]. [Online]. Available: http://www.newgeography.com/content/001680-how-texas-avoided-great-recession

[209] A. Hasenoehrl, "Montana's diverse economies: A regional labor market analysis," Tech. Rep., 2014.

[210] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[211] M. Sun and D. W. Wong, "Spatial aggregation as a means to improve data quality," in *Proceedings of the 13th International Conference on GeoComputation*, 2015.