A CYBERGIS ENVIRONMENT FOR SPATIOTEMPORAL ANALYSIS: A CASE STUDY OF CHINA MORTALITY DATA

BY

SU Y. HAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Geography
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Associate Professor Shaowen Wang

ABSTRACT

This thesis describes an online CyberGIS (cyberinfrastructure-based geographic information system) environment for a public health application. This CyberGIS application represents a new GIS application modality for protected access to public health data and related CyberGIS analytical and visualization services based on service-oriented architecture. A novel CyberGIS Open Service Application Programming Interface is employed to allow a number of users to execute multiple spatial analyses simultaneously on the backend powerful cyberinfrastructure. The user-centric interface of the CyberGIS application provides transparent access to cyberinfrastructure resources, which helps make the application scalable to the potential growth of the application users and pertinent public health researchers and users.

A case study is designed to examine spatial distribution characteristics of mortality rates in the mainland of China for the period of 1991 – 2000, using a suite of spatial data analysis methods integrated within the CyberGIS application that supports interactive exploratory analysis. The application demonstrates the novelty and utility of CyberGIS in this public health case study in the following four aspects. 1. Spatial pattern analyses combined with a set of visualization tools built on a CyberGIS infrastructure allow users to interactively explore the spatiotemporal distributions of heterogeneous data. 2. A highly interactive and easy-to-access user interface supporting temporal analysis allows users without in-depth GIS technical skills to experience an easy, efficient, and secure way to gain insights into the data. 3. Multiple users can access the application based on cloud computing support. 4. CyberGIS protects sensitive public health data based on a distributed database while enabling the dissemination of analytical insights based on the visualizations of spatial analyses.

ACKNOWLEDGMENTS

I am heartily thankful to my advisor – Dr. Shaowen Wang – and Mr. Yan Liu for their guidance, encouragement, and support during the research and preparation of this master's thesis. It would not have been possible to write this thesis without their thoughtfulness and patience in guiding my research. Dr. Shaowen Wang always gave me very strong motivation. His passion and positive energy was the greatest encouragement that enables me to finish this thesis. It has been a great honor to study under his direction. I also really appreciate Yan's technical support and very helpful discussions. I was so lucky to have Yan as my project manager. I believe that computational skills and knowledge learned from Yan will be a great asset for my future study.

I would like to thank my committee members, Dr. Sara Mclafferty and Dr. David Wilson. Dr. Sara Mclafferty directed me to study spatial analysis in public health. She also gave me mental and intellectual support during my master's study. Dr. David Wilson helped me to strengthen this thesis.

I am very grateful to Professor Gonghuan Yang – former deputy Director General of the Chinese Center for Disease Control and Prevention – for providing the mortality data of China used in this thesis. I would also like to acknowledge the following people who provided helpful comments on this thesis: Dr. Guofeng Cao, Eric Shook, and Yanli Zhao and all members in the CyberInfrastructure and Geospatial Information Laboratory.

I would like to express my gratitude and love to my family for supporting my education and great patience at all time. Lastly, I offer my regards and blessing to all of those who supported me in any respect during the completion of the thesis.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Online Web-based information environments are increasingly utilized in public health research and practice for information visualization and sharing through easy-to-access user interfaces. Within this broad context, geographic information systems (GIS) have become an important approach and tool in the domain of public health. Access to Web-based mapping and related spatial analysis, and usability and flexibility in making sense of spatial (i.e. geographically referenced) data in public health have drastically improved compared to conventional maps produced by desktop-based GIS software or printed atlases. Online GIS users can interact with dynamic maps through selecting and visualizing data on Web browsers (see e.g. http://www.earth.google.com). In addition, on the Internet many users can have shared access to specialized and customized GIS applications such as mortality and disease mapping, and disease surveillance systems (see e.g. http://www.who.int/research/en/).

Thus far, a number of studies on Web-based GIS for public health have focused on data sharing and interactive mapping (see Evans and Sabel 2012; MacEachren *et al*. 2008; Rop, Liu, and Wimberly, 2011; Toutant *et al*. 2011). Spatial analytical methods (broadly including spatial analysis and modeling) have been increasingly developed (Páez *et al.* 2010), and particularly these methods have been widely applied for supporting decision making in public health (Cromley and McLafferty 2011; Hanafi-Bojd *et al.* 2012; Watkins *et al.* 2012). On the other hand, a number of researchers have developed new methods based on desktop GIS software such as ArcGIS and combining data management with spatial and geo-visual analytical methods (Anselin *et al*. 2006; Guo 2008; Guo and Jin 2011). Open source GIS software such as GeoDa

(http://geodacenter.asu.edu/) and GRASS also has a variety of spatial analytical methods provided (http://grass.fbk.eu/). However, these GIS are designed based on the conventional computer-centric architecture (Wang 2010) in which it is often difficult to be integrated with powerful cyberinfrastructure resources. Cyberinfrastructure refers to integrated computing, information and communication technologies, and consists of computing systems, data, information resources, networks, digitally enabled-sensor, instruments, virtual organizations, and observatories, along with an interoperable suite of software services and tools (Atkins 2003; National Science Foundation 2007). If millions of users simultaneously run computationally intensive spatial analyses in the computer-centric architecture, any single powerful computer would be easily overloaded. For this particular reason, conventional GIS often lacks the capability of supporting computationally intensive spatial analysis and does not support any application environment in which a number of users can work collaboratively to solve geographical problems.

CyberGIS - new generation GIS based on cyberinfrastructure – aims to overcome the aforementioned limitations of conventional GIS by integrating cyberinfrastructure, GIS, and spatial analysis capabilities (Wang 2010). CyberGIS enables spatial analytical capabilities to be integrated with Web-based mapping and powerful cyberinfrastructure resources such as high-performance computers and cloud computing resources. Another important aspect of CyberGIS is its focus on sharing of digital services among a large number of users for collaborative geographic problem solving and decision-making.

The primary purpose of this thesis research is to establish a CyberGIS environment with a highly interactive and multi-user online interface empowered by a suite of visual and analytical CyberGIS services. A case study is conducted to demonstrate how the CyberGIS environment

can support public health researchers to explore and analyze spatiotemporal mortality data through on-demand online maps, charts and analytical services. The challenge of organizing, visualizing, and analyzing spatiotemporal mortality data and high-dimensional socioeconomic and demographic data is tackled by integrating visualization tools, interactive mapping and cyberinfrastructure-based computation into the CyberGIS environment based on service-oriented architecture.

In the case study, two spatial analysis methods are adapted to allow users to gain further insights beyond data visualization and interactive mapping. The motivation for the use of spatial analysis is to represent the mortality rates of non-sampled counties by using sampled counties in which the mortality rates are known, and to identify regions with relatively high mortality rates compared to mortality rates in other regions. One method is self-organizing map (SOM), widely used for exploratory knowledge discovery (Agarwal and Skupin 2008). The other is a straightforward spatial classification (SSC) method, the development of which is based on SOM.

These analyses focus on deriving the patterns of the mortality rates across all counties from sampled counties by utilizing two available datasets: (1) China national mortality data that are sampled in 145 selected counties among 2420 counties of the mainland of China from 1991 to 2000; (2) high-dimensional socioeconomic and demographic data from the 1990 census collected at county level (see section 4.1 for details). Because the available data that can be used to estimate the mortality rates of non-sampled counties are limited, it would not be possible to construct a reliable model to predict the mortality rates of non-sampled counties. Therefore, instead of using a statistical approach to the prediction of the mortality values of non-sampled locations, our analyses (SOM and SSC) are performed based on the principles of exploratory data analysis (EDA) (see section 4.3 for details). EDA approach can provide useful hints that are

needed to estimate the mortality rates of non-sampled counties. Then, the estimated mortality rates help to reveal the general patterns of mortality rates across all counties in China. In addition, understanding the general patterns would be able to help create a hypothesis to design statistical models for the next level of study.

To achieve the objective of revealing the patterns of mortality rates by using aforementioned available datasets, our approach consists of two parts—classification and association. The first part is to group socioeconomically and demographically similar counties together. Among the counties within the same class defined at the first part, the second part is to select the representative sampled county and associate the mortality rate of the representative county to the mortality rates of non-sampled counties.

For the first part, it is challenging to precisely measure the similarity of high-dimensional socioeconomic and demographic data among counties. The SOM is applied to address this challenge by generating multiple classes to which socioeconomically and demographically similar counties belong. Regarding the second part for achieving the association, each class should include only one sampled county and one or more non-sampled counties. Through this association process, among the counties belonging to the same class, the mortality rate of the sampled county is assigned to the mortality rates of the non-sampled counties. A problem, however, is that the classification of SOM may include classes to which multiple sampled counties belong or it may not include any sampled county. In the former case, it is often difficult to identify the representative sampled county. In the latter case, there is no representative mortality rate of the sampled county that can be associated to the mortality rates of the non-sampled counties. Consequently, given the SOM classification, the mortality rates of some non-

sampled counties are not feasible to be associated with the morality rates of the sampled counties. Therefore, SSC is developed to achieve the desirable association.

SOM provides a foundation for the development of SSC and also has its own value as a knowledge discovery tool. SSC is built on SOM with respect to the measurement of the similarity of high-dimensional socioeconomic and demographic data, but uses a different classification method that is distinguished from the SOM's classification. Though SOM itself does not directly contribute to the revelation of the patterns of the mortality rates of all counties, it results into various views of graphs and maps that can be useful for further analysis.

In this thesis research, by integrating the aforementioned spatial analyses, health-related data, and a highly interactive user interface in the CyberGIS environment, users experience dramatically reduced turnaround time required for performing spatial analyses. In addition, problems of conventional GIS regarding the limitations in its computational scalability are resolved by employing cyberinfrastructure, access to which CyberGIS makes easy and intuitive by developers and users. The research is intended to address the following three questions. 1. Given the limited number of samples from the available datasets, how the CyberGIS application can facilitate the analysis and revelation of spatiotemporal patterns of the China national mortality rates? 2. Are there any particular spatiotemporal patterns of the mortality rates of certain diseases at the national scale? And 3. How can the CyberGIS environment benefit public health researchers and practitioners?

The rest of this thesis is organized as follows. The background section reviews related work on Web-based GIS in public health and spatial analysis in medical geography. The section on the CyberGIS environment describes the architecture, interface design, and components of the CyberGIS environment. The case study addresses specific methods and focuses on findings

resulted from spatial analyses within the CyberGIS application. Finally, the summary and conclusions section summarize the findings of this thesis, and address the significance of the findings.

CHAPTER 2

BACKGROUND

In this chapter, the first-part discussion is focused on literature related to Web-based GIS applications in public health and addresses needs for CyberGIS. Then, it reviews the use of self-organizing map (SOM) in health geography research. SOM is adapted for analysis of sparsely sampled mortality data of China, which will be further detailed in the case study of this thesis.

2.1 WEB-BASED GIS IN PUBLIC HEALTH

GIS is widely used for managing and analyzing public health data and revealing spatial patterns that may be difficult or impossible to be discovered. However, using conventional GIS requires mastering sophisticated skills in order to fully utilize necessary techniques and methods to generate desirable results and maps. On the other hand, public health researchers or policy makers have increasingly recognized the need for sharing geospatial data and using GIS for problem solving and decision-making. Web-based GIS emerged to enable easy access to geospatial information derived from public health data to broad users including those without in-depth GIS expertise (Cromley 2003; Croner 2004). While a number of Web-based applications have recently been developed to resolve various public health issues, including for example disease surveillance (Robertson and Nelson 2010), GIS applications for geospatial knowledge discovery and decision making in public health remain to have tremendous needs and require major advances to be made in GIS and spatial analysis tools (Supak *et al*. 2012).

2.1.1 Commercial off-the-Shelf Web-based GIS

Conventional Web GIS implementations are often led to isolated, standalone, monolithic and proprietary systems (Anderson and Moreno-Sanchez 2003), which focus on data and tools implemented with the client-server architecture (Rinner 2011). For example, Blanton *et al*. (2006)

developed a spatial database and real-time Internet mapping tool for rabies surveillance in the United States by utilizing commercial Web GIS software—i.e. ArcIMS software. By using the same software, Yang *et al*. (2007) built a spatial decision support system for epidemic disease prevention.

Meanwhile, Kamadjeu and Tolentino (2006) suggested that proprietary GIS technologies are a limiting factor in the adoption of GIS in public health organizations that in many cases lack resources such as GIS hardware, software, and budgets for acquiring necessary technical expertise. When such organizations try to use commercial Web GIS software to exchange geospatial data and deliver functions of spatial and statistical analysis, they often come across the following issues: (1) such GIS does not offer out-of-the-box spatial analysis functionality to support customized analyses; (2) it requires long-term commitments to cope with software evolution; (3) it requires that some of their IT personnel become specialists in the software operation and maintenance; and (4) it is costly to integrate with existing IT infrastructure. For these reasons, commercial off-the-shelf GIS are often beyond the reach of resource-constrained public health organizations (Anderson and Moreno-Sanchez 2003).

2.1.2 Open Source Web-based GIS

Open source GIS has emerged to overcome some of the aforementioned issues. Open source technologies can be shared with anyone and allow resource-constrained public health agencies to use Web resources with low development cost (Yi *et al*. 2008). Personnel with general IT background can learn open source technologies. In addition, open source GIS is freely extensible with functionalities that are not available in commercial software, and compatible with existing IT infrastructure including personnel skills, software, and applications (Anderson

and Moreno-Sanchez 2003). Therefore, open source GIS software may be desirable for resource-constrained public health agencies (Richards *et al*. 1999).

Recently there are numerous health-related Web-based GIS applications which address the potential of open source software in terms of its interoperability and scalability (Evans and Sabel 2012; Kamadjeu and Tolentino (2006); Maclachlan *et al*. 2007; Pirotti, Guarnieri, and Vettore 2011; Toutant *et al* 2011;Vanmeulebrouk *et al*. 2008). Boulos and Honda (2006) pointed out that open source Web-based GIS software systems have reached a stage of robustness and stability rivaling that of commercial Web-based GIS and provided instructions on how to publish their own health maps with Web Map Service (WMS) support. Moreno-Sanchez *et al.* (2007) emphasized the potential of open source software by utilizing distributed raster images that are generated as map layers. MacEachren *et al*. (2008) presented the use of Web Feature Service (WFS) to support highly interactive and dynamic geo-visualization tools in the development of Web-based GIS-enabled cancer atlas across the state of Pennsylvania.

2.1.3 Service-Oriented Web-based GIS

Architecture of Web-based GIS now is evolving to open, distributed, service-oriented environments where data can be transparently exchanged among components offering specific geo-processing functionalities (Dangermond 2002). In this architecture the Web is used for delivering not only data, but also spatial analysis methods or geo-processing tools that can be wrapped in interoperable software components (Anderson and Moreno-Sanchez 2003). These components can be integrated together to build comprehensive services or applications (Hecht 2002). Within this context, Croner (2003) demonstrated the potential of Web resources for public health decision-making and the integration of distributed spatial data applications into cyberinfrastructure including supercomputing, and data transfer and mining technologies. Goa *et*

al. (2008) presented an interoperable service-oriented architecture for online mapping of spatiotemporal disease information, and augured that such an infrastructure enhances efficiency and effectiveness of public health monitoring. Tiwari and Rushton (2010) developed an environmental health surveillance system (EHSS) that serves to (1) visualize the spatial patterns of diseases while preserving privacy, and (2) automatically link environmental data, environmental models, and geo-processing functionality to estimate individual exposures to environmental contaminants. The authors presented a modular, Web-based spatial analysis system that uses spatial analysis methods and services delivered over computer networks. Rinner *et al*. (2011) proposed service-oriented architecture that integrates publicly accessible map services with protected public health data layers to investigate injury rates and demographic factors through a spatial lens. Supak *et al* (2012) created a flexible framework for public health monitoring while focusing on the flexibility and scalability of the framework that allows the system to be customizable, modular, portable, and easily configurable to support additional research and education initiatives.

2.2 NEEDS FOR CYBERGIS IN PUBLIC HEALTH

Recently, Web-based public health applications are omnipresent. Unfortunately, many of these applications are built as closed systems where it is difficult to use spatial analysis tools or geo-processing functionalities of other services and integrate with cyberinfrastructure resources. In addition, in a closed system a number of users who can simultaneously run spatial analyses or geo-processing functionalities on the Web are often limited. Wang (2010) showed that those systems have computer-centric architecture where hardware and operating systems are treated as center and databases and application software are built as peripherals. Consequently, Web-based GIS built in this architecture is often limited to enable collaborative problem solving involving

multiple users' shared access to GIS services, and is also lacking support for computationally intensive spatial analysis.

CyberGIS addresses aforementioned issues so that a large number of users can be allowed to perform spatial analyses collaboratively and simultaneously, and exchange services on cyberinfrastructure including resources such as high-performance computers, clouds, remote visualization systems, and capabilities of knowledge management and virtual organization support functions. The holistic approach of CyberGIS is to interlink both application-driven and user-centered functionalities of cyberinfrastructure, GIS, and spatial analysis.

One important capability of CyberGIS is to enable computationally intensive spatial analyses running on powerful cyberinfrastructure, while user-centered interfaces enhance the accessibility and usability of CyberGIS applications. Through intuitive interfaces, users can easily perform spatial analyses without going through a steep learning curve that is often required in desktop-based GIS. Another important aspect of CyberGIS is its service-oriented and open framework. For example, application developers can use OpenLayers as a client-side mapping tool, GeoServer for managing layers and styles of visualization, Apache Web server for message handling and data retrieval, Ext JS or Yahoo User Interface (YUI) Library for visualization tools and layouts, and cloud computing or supercomputing to scale up the capacity of spatial analyses, and remote database for protecting sensitive data. CyberGIS synthesizes all of these digital resources and services, and achieves a cohesive single environment. Therefore, by utilizing CyberGIS for supporting knowledge discovery and decision making in public health, a wide range of users such as public health professionals, researchers, policy makers, and general public, can not only share geospatial information, but also work jointly on solving complex geographic problems.

2.3 SPATIAL ANALYSIS – SELF-ORGANIZING MAP

Self-organizing map (SOM) has been widely used as data clustering, classification, visualization, dimension reduction and pattern recognition methods. Several researchers used SOM to reveal and visualize structures of high-dimensional public health data and explore the relationships among data attributes (Mehmood *et al.* 2011; Ki *et al*. 2011; Koua and Kraak 2004; Törönen *et al*. 1999; Valkonen *et al*. 2002). Also, SOM has been utilized to analyze spatial patterns in public health research. Oyana *et al*. (2005) suggested the potential use of SOM for analyzing spatial data in biomedical domains. They applied SOM to explore the patterns of adult asthma patient data and revealed that asthma is more prevalent in areas that are close to major roadways and pollution sources. Basara *et al*. (2008) applied SOM to classify high-dimensional environmental variables. They found that there is a significant relationship between SOM classifications and the geographic distribution of diseases and, also revealed that the environment is correlated with the distribution of both chronic and infectious diseases. Zhang *et al*. (2009) identified the groups of geographical areas that share similar epidemiological data attributes. Previous work suggests that SOM as an exploratory data analysis method is effective in public health research for analyzing high-dimensional data including for example health outcomes, socioeconomic and demographic variables, and physical environments.

CHAPTER 3

CYBERGIS ENVIRONMENT

This chapter describes the architecture, interface design, and components of a CyberGIS environment tailored to the spatiotemporal morality data and related public health application.

3.1. ARCHITECTURE OF CYBERGIS ENVIRONMENT FOR PUBLIC HEALTH

The CyberGIS application is developed based on the principles of service-oriented architecture, and makes use of open source software for flexible customization of the technologies involved (Figure 1). The entire application includes PostGIS (remote database where sensitive public health data are stored), GeoServer for rendering and configuring the visualization of map layers, Web server for user interaction handling and data retrieval, and OpenLayers and Yahoo User Interface (YUI) Library for programming client-side user interfaces. In addition, the spatial analyses – self-organizing map and the straightforward spatial classification – are deployed within a cloud infrastructure of the CyberInfrastructure and Geospatial Information Laboratory (see chapter4 for the detailed description of the spatial analyses). Service-oriented architecture and cloud computing combined with open source software assure the scalability and interoperability of the CyberGIS application. The user-centric interface of the CyberGIS application provides transparent access to CI resources, which help make the application scalable to the potential growth of the application users.

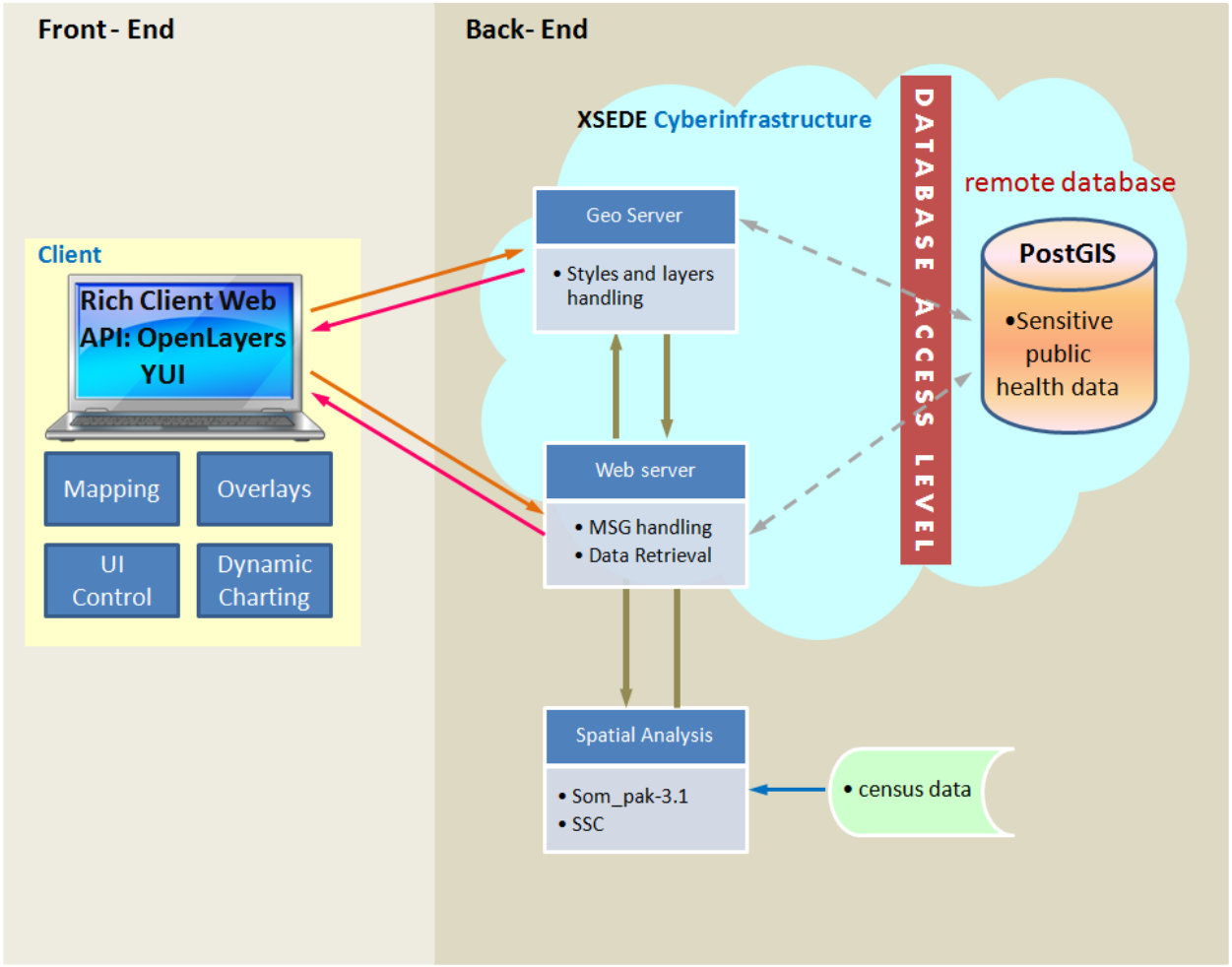


Figure 1

Architecture of the CyberGIS application

• XSEDE stands for the Extreme Science and Engineering Discovery Environment, an integrated CI environment supported by the National Science Foundation (see https://www.xsede.org/).

• Census data: 40 socioeconomic and demographic attributes from the 1990 census of China

• Sensitive public health data: mortality data from 1991 to 2000 from the Chinese Center for Disease Control and Prevention

• Som_pak-3.1: self-organizing map software package

• SSC: straightforward spatial classification

Spatial analysis based on cloud computing provides on-demand access to computational resources (Figure 1). In the client-server architecture of GIS, when an application becomes

popular and has a large number of users, a single server may become overloaded. However, in our user-centered environment, the analyses are deployed to many virtual machines in the cloud, which enables serving massive users in a scalable way. A basic idea of cloud computing is that computing capacity available to users is elastic. The computing resources could be expanded or also shrunken based on users' dynamic requests to run spatial analyses. Through this cloud computing approach, a large number of users can run numerous analyses simultaneously without concerning about how much computing power is available to them. Therefore, dispatching spatial analyses to cloud infrastructure can help scale up the capacity of spatial analyses. For the integration of the user environment and back-end cloud infrastructure in which the spatial analyses are computed, an Open Service API is used. CyberGIS Open Service API (created by the CyberInfrastructure and Geospatial Information Laboratory at the University of Illinois at Urbana Champaign) bridges between application clients and back-end CI resources such as cloud computing. This API streamlines the integration of spatial analyses with CI. In addition, in the future other researchers who want to integrate the existing spatial analyses will be able to integrate them in their applications through the Open Service API.

The integration of remote database access and control can contribute to the dissemination of sensitive geospatial information derived from public health data while preserving the confidentiality of health records (Figure 1). Health organizations are often careful about releasing public health data that can possibly lead to the identification of individuals (Croner 2003). Consequently, researchers have developed a number of methods for geographically masking (i.e. modifying) the locations of individuals containing health records (Armstrong *et al.* 1999). The capability of remote database access and control allows a wide range of user privileges to be specified for flexible protection of the security of public health records. Since the

remote access to database is feasible, public health agencies do not need to release their sensitive health records out of their own database while permitting spatial analyses on their data. Spatial analyses generate maps visualizing trend and distributional patterns of diseases rather than presenting individual records. Furthermore, results of spatial analyses often allow users to have more advanced insights than visualizing raw data. Then, visualizations of spatial analysis results can be widely shared while ensuring the prevention of the disclosure of original health records. In aggregate, the CyberGIS application allows for different rights of accessing the database for different users. For example, general public can query to see the distribution of mortality rates of stomach cancer. Doctors can have access to sensitive data (e.g. sexually transmitted disease) that contain identifiable personal information.

3.2 INTERFACE DESIGN AND COMPONENTS

With the advance of Web 2.0 technologies, CyberGIS applications can have user-friendly interfaces and, thus allow for easy access by those who do not possess GIS technical expertise. Target users for our CyberGIS application may include decision makers in public health organizations, public health researchers, and even general public users. The user interfaces consist of two parts: “Temporal Chart” (Figure 2) and “Spatial Chart” (Figure 3). Points represent the locations of stations where mortality data were sampled (description of the data and the study area are detailed in the section 4.1). Each part provides highly dynamic and interactive charting tools. Spatial analyses are arranged in the Spatial Chart part with a bar graph representing spatial distributions of mortality rates.

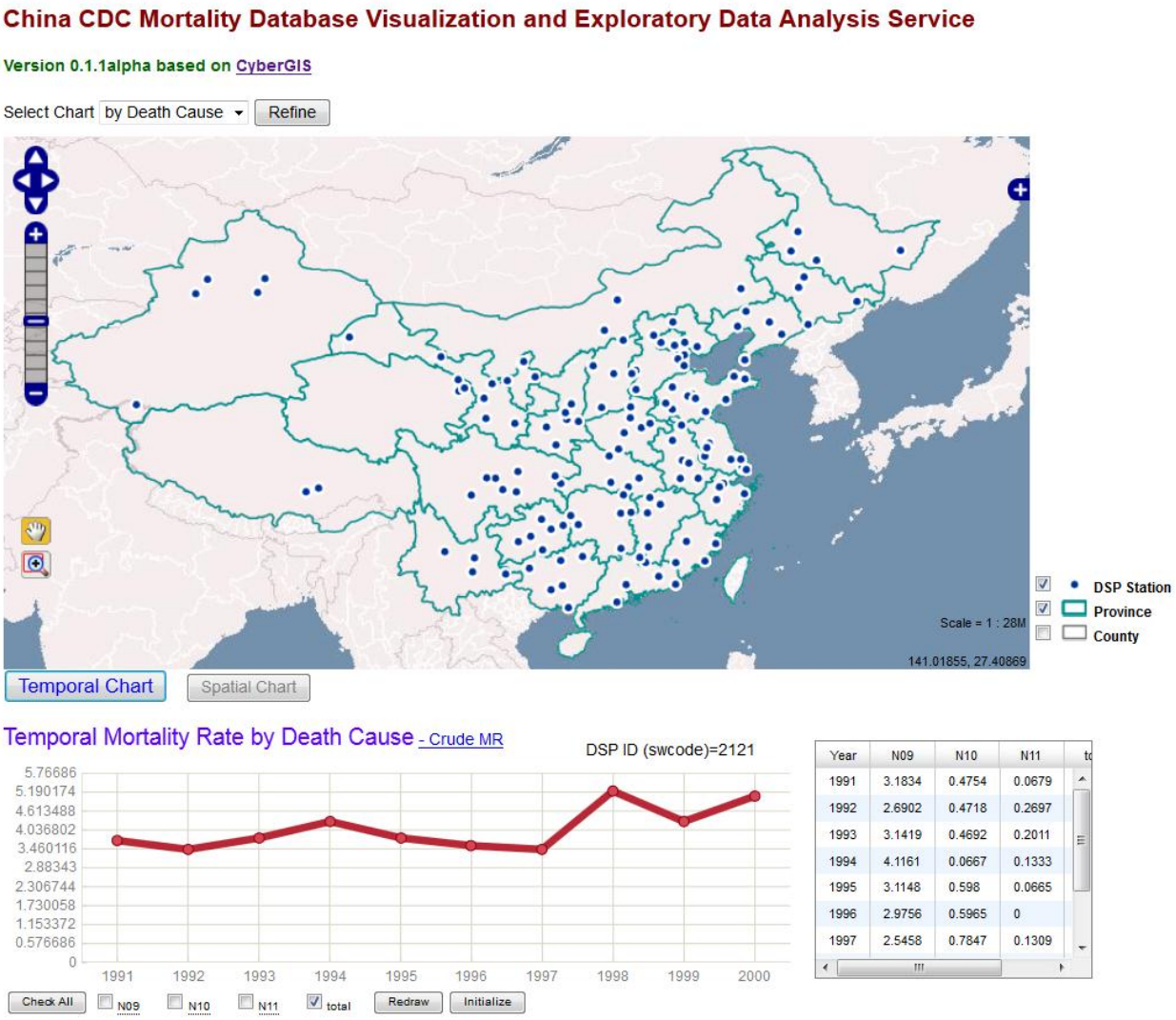


Figure 2

The interface of "Temporal Chart" of the CyberGIS environment: Buttons and check boxes below the line graph allow users to compare and contrast the temporal changes of the mortality rates of each disease (N09, N10 and N11) among the selected population (males in the range of ages between 40 to 85). Currently only the total mortality rate of all three diseases is shown. Various views of this "Temporal Chart" are available in Figure 7, 8 and 9. A table view is available on the right side of the chart.

N09 (malignant neoplasm of digestive organs and peritoneum)
N10 (malignant neoplasm of respiratory and intrathoracic organs)
N11 (malignant neoplasm of bone, connective tissue, skin, and breast)

Figure 3

"Spatial Chart": a different view of this chart is available in Figure 10.

3.2.1 Data Selection

The interfaces allow users to explore the temporal and distributional differences and similarities of mortality rates as a whole and across categories of gender, age, and death causes. For the calculation of mortality rates, there are options for users to select input parameters on the top of the map (Figure 4): gender (male and female), age (0, 1, 5...80, 85, and >100), and death causes that are represented based on the Chinese Classification of Diseases (CCD).

Figure 4

On the interface, the user can select input parameters to visualize mortality rates in maps and charts and perform spatial analyses. There are three major categorizations—i.e. gender, age, and death causes. After this major categorization is chosen, the user can have more options to refine input parameters by clicking the "Refine" button.

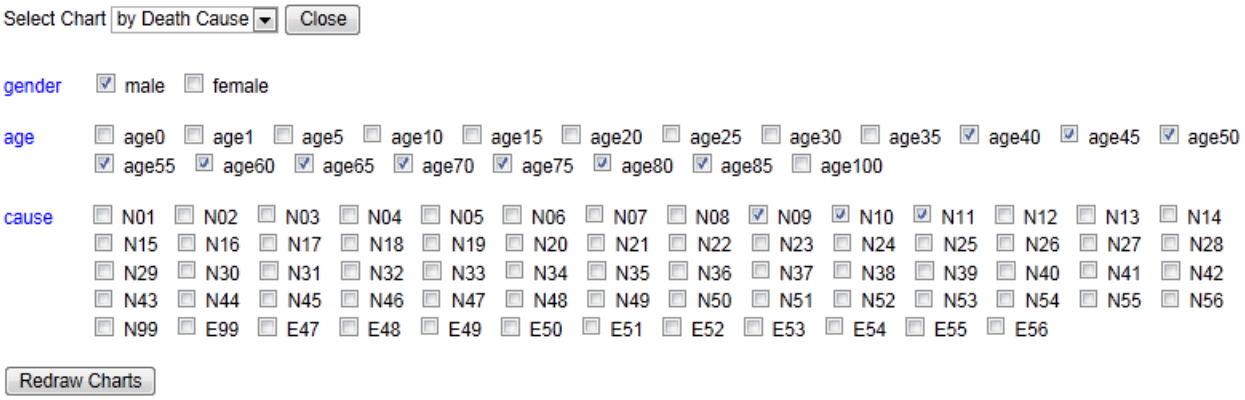


Figure 5

A user's selection of input parameters

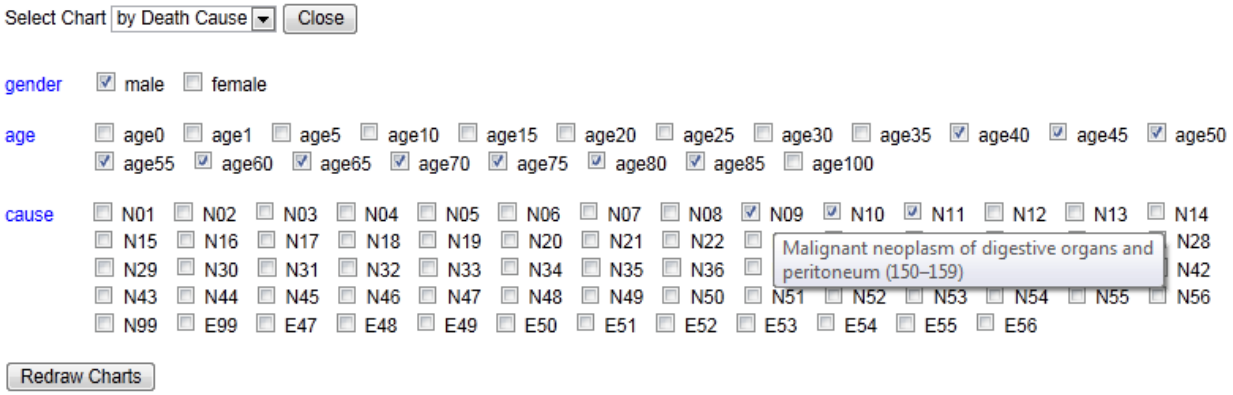


Figure 6

A tooltip to show what N09 represents

The CCD code starting with N means internal causes of death and the code starting with E represents external causes of death. When the user hovers over each code, a tooltip containing its corresponding name of a specific cause of death and its ICD-9 code pops up. For example, N09 indicates the malignant neoplasm of digestive organs and peritoneum, and 150-159 represents corresponding ICD-9 code for N09 (Figure 6). To define input parameters, there are a major categorization (Figure 4) and further refinement of input parameters (Figure 5). The user can choose one among major categories—i.e. gender, age, and death causes. When the "Refine" button is clicked (Figure 4), the user has additional options to refine input parameters (Figure 5). For example, the user selects death causes among the major categories, and then selects male, aged 40 to 85 years old, and N09 (malignant neoplasm of digestive organs and peritoneum), N10 (malignant neoplasm of respiratory and intrathoracic organs), and N11(malignant neoplasm of bone, connective tissue, skin, and breast) among death causes. Then, the mortality rates of the chosen diseases among the male population in the range of ages between 40 and 85 are calculated and visualized using both the "Temporal Chart" (Figure 2) and "Spatial Chart" (Figure 3). Since the user selects death causes in the major categorization, the temporal chart gives the user options to temporally compare and contrast the mortality rates of each disease (N09, N10 and N11) among the selected population (males in the range of ages between 45 to 80) (Figure 2).

3.2.2 Temporal Changes of Mortality Rates

Crude mortality rate is used to show their temporal change. It is the total number of deaths to residents in a specified geographic area (county) divided by the total population of the same geographic area (county) and multiplied by 1000. The default "Temporal Chart" first displays the temporal change of a total mortality rate of all of the sampled counties that are

represented as points on the map (Figure 2). As the user specifies input parameters (e.g. Figure 5), the temporal changes of mortality rates specific to age, sex, or certain causes of death are visualized. Also, the temporal changes of the mortality rates of a specific location are visualized on the line chart, whenever the user clicks one of the point locations on the map. For example, with the selection of input parameters as described above (Figure 5), once the user clicks a point on the map, a line graph is created to show the temporal change of mortality rates of the clicked location (Figure 2). The chart first displays a total mortality rate of the three selected diseases among the male population in the range of ages between 40 and 85, and has options to visualize N09, N10, and N11 (see the unchecked check boxes in Figure 2). After the user clicks the "Check All" button, all of the three lines are added to visualize the temporal changes of mortality rates of each disease (Figure 7). When the user hovers over each line, a tooltip pops up and tells what the line represents (Figure 8).

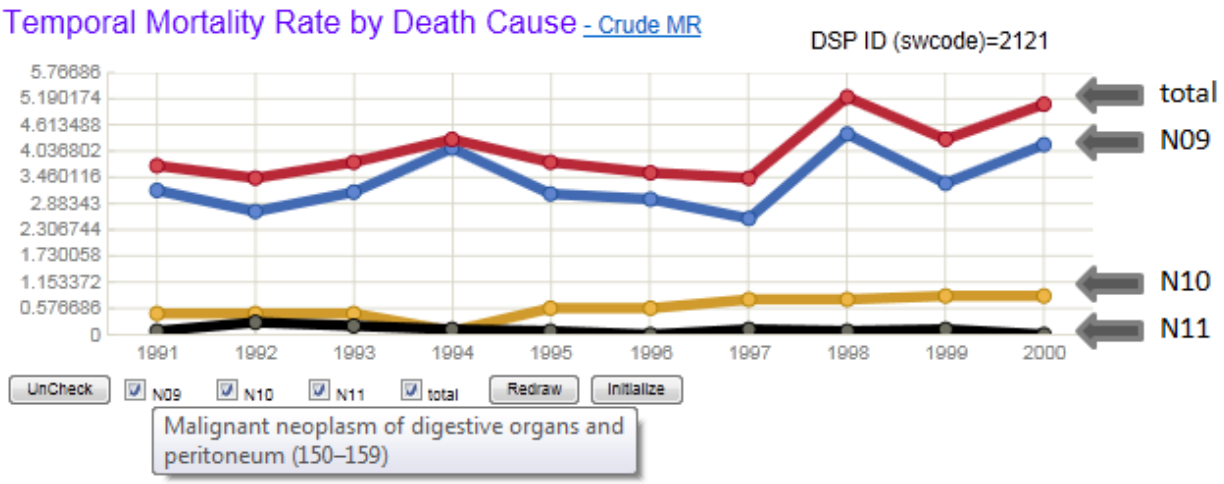


Figure 7

This graph is captured from the page of "Temporal Chart" in Figure 2. The image is captured when the user's mouse hovers over N09 (Note a tooltip).
red: total
blue: N09 (malignant neoplasm of digestive organs and peritoneum)
yellow: N10 (malignant neoplasm of respiratory and intrathoracic organs)
black: N11 (malignant neoplasm of bone, connective tissue, skin, and breast)

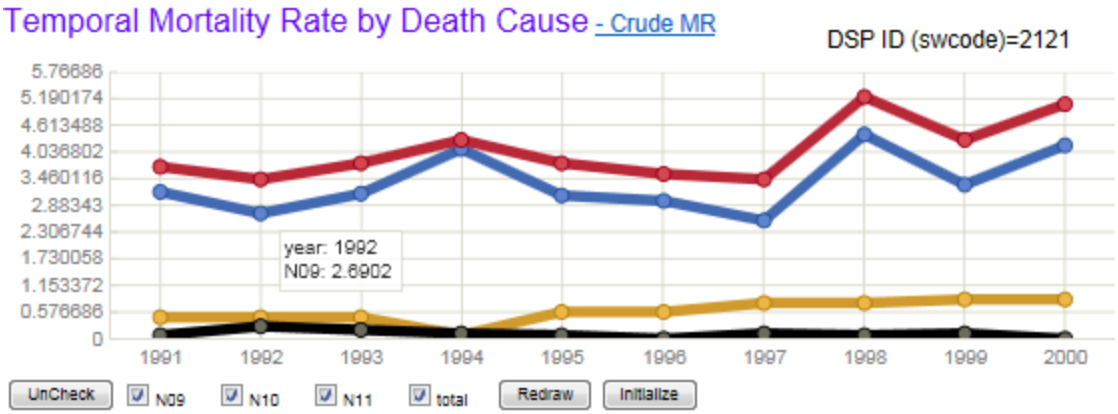


Figure 8

A tooltip pops up when the user mouse hovers over the second circle of a blue line.

In addition, the "Temporal Chart" allows users to zoom in certain parts of the line graph. The line graph of Figure 7 represents that the mortality rates of N09 are always much higher than the mortality rates of N10 and N11 for the 10-year period from 1991 to 2000. However, the temporal changes of the mortality rates of N10 and N11 cannot be obviously shown. Especially, a line of mortality rates of N11 is almost crammed at the bottom of the graph because of the big difference between the mortality rates of N09 and N11. In this case, the user can uncheck the checkboxes of the total and N09 and hit the "Redraw" button (see Figure 9).

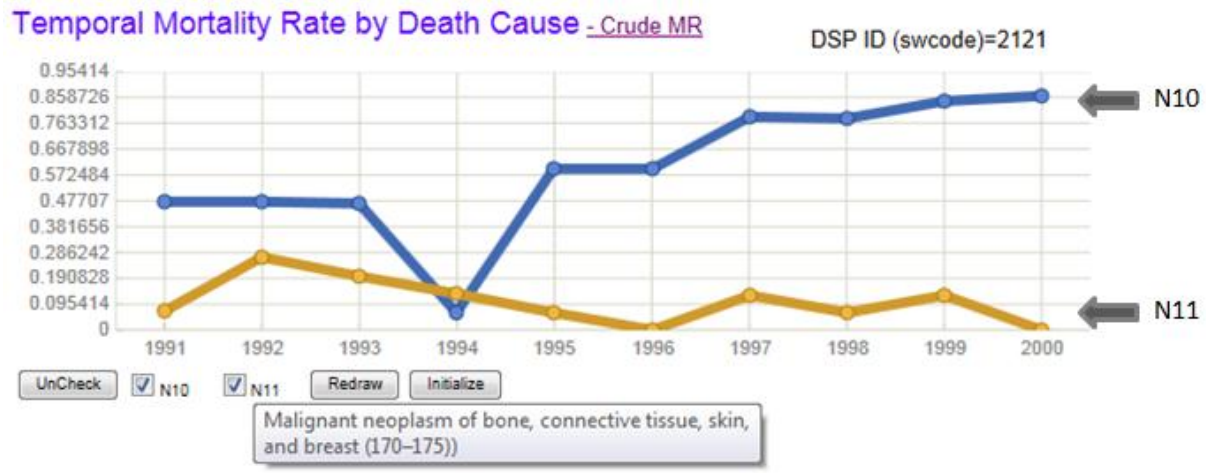


Figure 9

This chart is redrawn from the chart of Figure 7.
blue: N10 (malignant neoplasm of respiratory and intrathoracic organs)
yellow: N11 (malignant neoplasm of bone, connective tissue, skin, and breast)

Then, the graph zooms in the range where the mortality rates of N10 and N11 are visualized. Before the user redraws the graph, the maximum mortality rate of the *y*-axis is 5.77 (see Figure 7). In contrast, in the redrawn graph, the maximum mortality rate of *y*-axis is 0.95 (see Figure 9). Consequently, the redrawn graph (Figure 9) shows temporal variations of mortality rates of N10 and N11 more clearly than the prior graph (Figure 7). The mortality rate of N11 has a decreasing trend, which cannot be examined in the graph of Figure 7. When the user hits the "Initialize" button, the graph goes back to its initial view (Figure 2). In this way, the user can selectively visualize the line representing a temporal change of mortality rates, and also compare and contrast multiple line series in the graph.

3.2.3 Distributional Patterns of Mortality Rates

In this section, all the mortality rates are age-adjusted based on a direct age standardization method (Ahmad *et al*. 2001). While the crude mortality rate is useful for determining the magnitude of health status of a geographic area, it is not appropriate for the comparison of the mortality rates of populations in different geographic areas (see Curtin, Klein, and National Center for Health Statistics (US) 1995). When the crude mortality rate is used to compare the mortality rates of populations belonging to groups having the different age composition, it brings the following issue. Because the mortality occurs at infant and old-age classes, the population with a large number of infants and old people would have a higher mortality rate than the population with a small number of groups of infants and old people. In order to eliminate the effect of the difference of the age composition in calculating mortality

rates, the crude death rate should be adjusted for the differences in age composition between the population of interest and a standard population. The formula is given as follows:

Direct Standardization of (Age Adjusted) Mortality Rate = $[\text{Sum}_{\text{age groups}} (M_{ar} P_{as})]/P_s \times 1000$

M_{ar} is the age-specific mortality rate for the region.

P_{as} is the number of people in the age group in the standard population.

P_s is the total standard population.

"Spatial Chart" provides comparative views of the mortality rates across province and counties every year from 1991 to 2000 (Figure 10). For example, after the user's selection of input parameters as defined in Figure 5, the chart represents the total mortality rates of all N09, N10, and N11 among the male population in the range between ages 40 and 85 at the province level and at the county level in the year 2000.

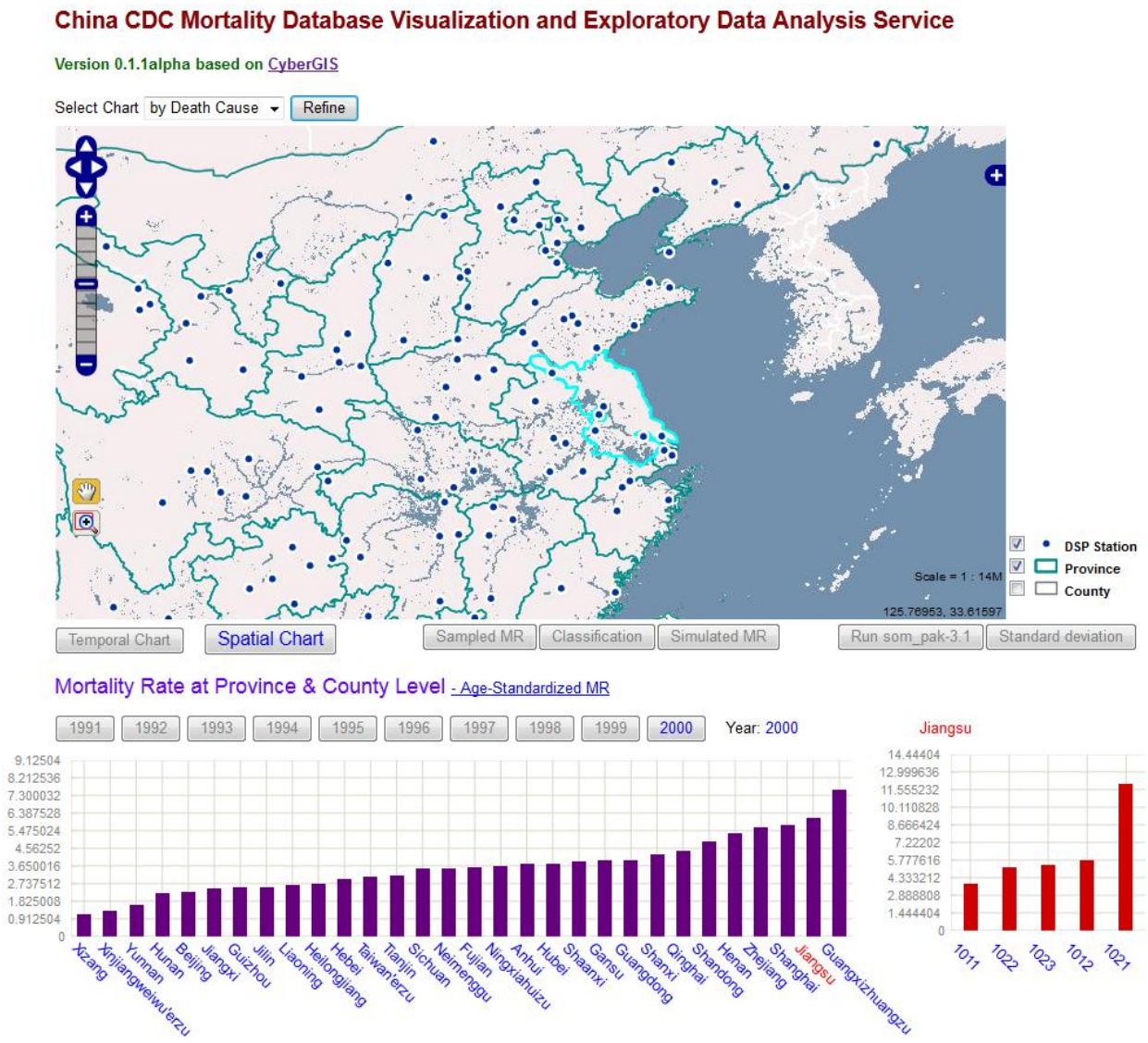


Figure 10

The user clicks Jiangsu province. Then, the Jiangsu province is highlighted on the map and also another chart is drawn on the right side. The chart represents the mortality rate of each five sampled counties within the Jiangsu province.

For the province-level visualization, the chart visualizes an average of the chosen mortality rates of all the stations within each province. The right-most bar of the chart represents the Guangxi province having the highest average mortality rate of all the stations within every province. In contrast, the left-most bar represents the Xizang province representing the lowest average

mortality rate of all the stations within every province. In addition, the chart allows the user to click one of province names on labels. Then, another chart comes out on the right side and shows the mortality rate of each sampled county of the clicked province. For instance, in Figure 10 Jiangsu province is clicked by the user. Then, the Jiangsu province is highlighted on the map and also another chart is drawn on the right side. The chart represents the mortality rate of each five sampled counties within the Jiangsu province. In this fashion, linking between a map and the charting tool enables users to drill-down easily from a province level overview to more specific information about places of interest.

Users can run spatial analyses on the page of "Spatial Chart" - e.g. Figure (the description of the spatial analyses is provided in the section 4.3). Before users run spatial analyses, they can visually examine the mortality rates of sampled counties that are one of inputs of spatial analyses. For example, after "Sampled MR" button and year "1992" button are clicked, the image of Figure 10 is shown on the map area. When "Classification" button is clicked on the interface (e.g. Figure 10), the map shows the classification result of the straightforward spatial classification method (e.g. Figure 36). "Simulated MR" button is to show the distributional patterns of mortality rates in the entire counties including sampled counties (e.g. Figure 38). Both buttons: "Run-som_pak-3.1" and "Standard deviation" allow users to run self-organizing map and discover knowledge based on the various visualizations of the results (e.g. Figure 18 and 22). SOM_PAK3.1 is standalone software created by the Neural Networks Research Center at the Helsinki University of Technology (Kohonen *et al*. 1996). The source code, written in C, is freely available for non-commercial uses. When users click a button "Run-som_pak-3.1," the source code is executed within the CyberGIS environment.

3.2.4 Combined Uses of Spatial Analyses and Charting Tools

Spatial analyses and charting tools can be used synergistically. For example, mortality rates are defined as shown in Figure 5. Then, the user visualizes defined mortality rates by using charting tools and distributional patterns of the mortality rates produced by performing spatial analyses (e.g. Figure 11, 12 and 13).

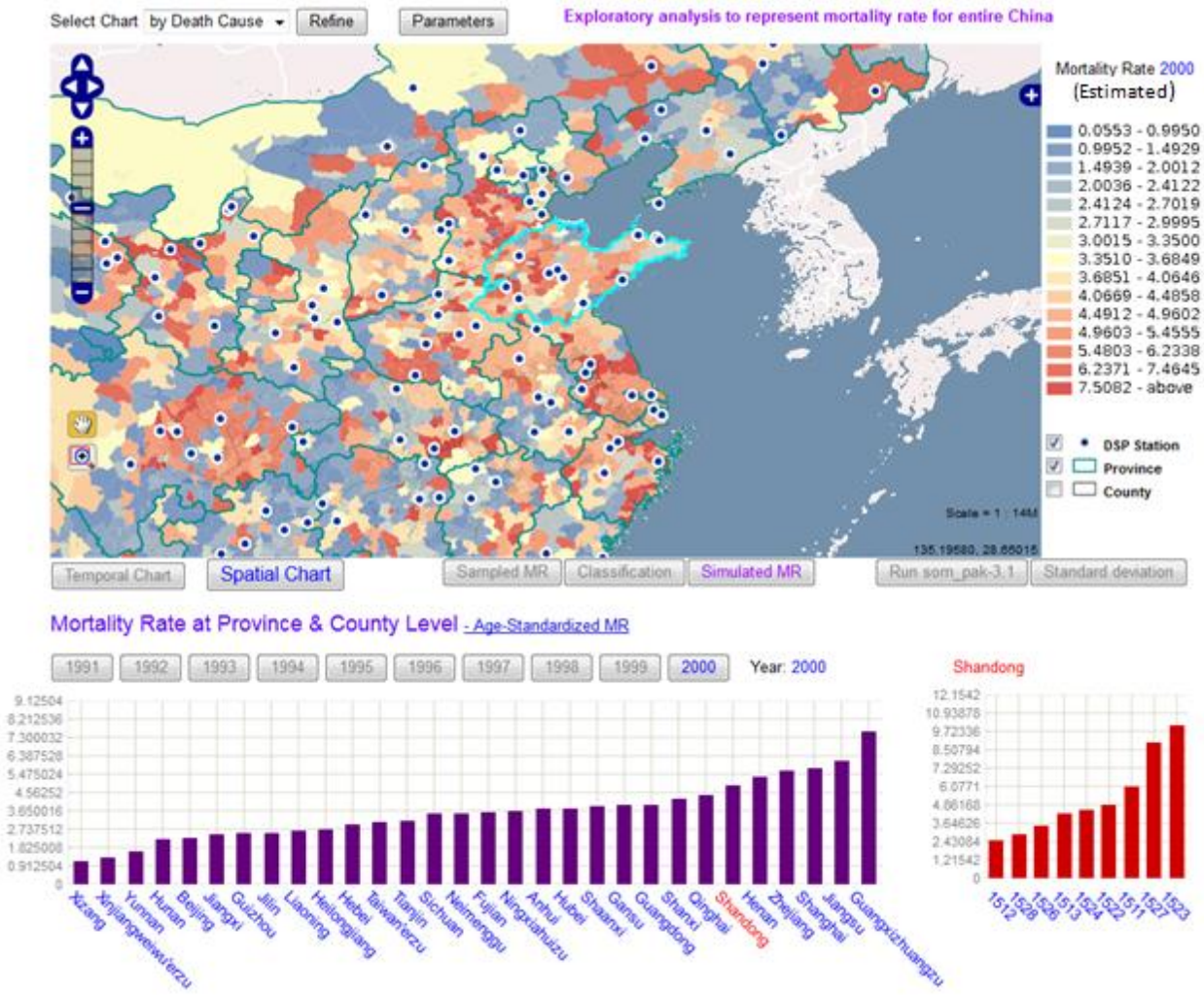


Figure 11 The image on the map (above) is after the user zooms in Shandong province. The image of entire China is shown below. The graph on the right shows the mortality rate of each sampled county within the Shandong province.

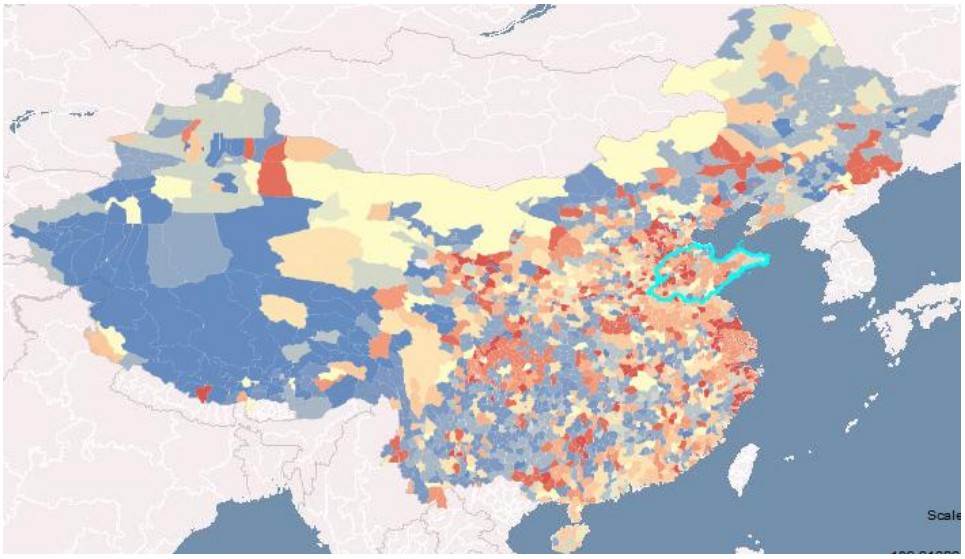


Figure 12

Shandong province (highlighted on the map) shows the relatively high mortality rates of the selected diseases (N09, N10 and N11) compared to other parts, especially the west parts of China. Counties in Shandong province are mostly colored in orange or red, which means the potentially high mortality rates of the selected diseases.

As the users examine spatial distributional patterns of mortality rates of China, they are probably interested in focusing on the province level showing the high mortality rates as represented in red color on the maps. Then, the users may click a province showing the high mortality rates on the chart and examine the county level mortality rates within the selected province while simultaneously examining the red colored regions on the map. For example, Shandong province is focused since most counties in the province are represented in oranges or red indicating the relatively high mortality rates (see the highlighted province on the map in Figure 11). In this case, the user might also be interested in investigating the temporal changes of mortality rates of each sampling station within the Shandong province. Figure 13 shows that one of the stations in the selected Shandong province is clicked and all of the mortality rates of N09, N10, and N11 indicate increasing trends during the ten-year period from 1991 to 2000.

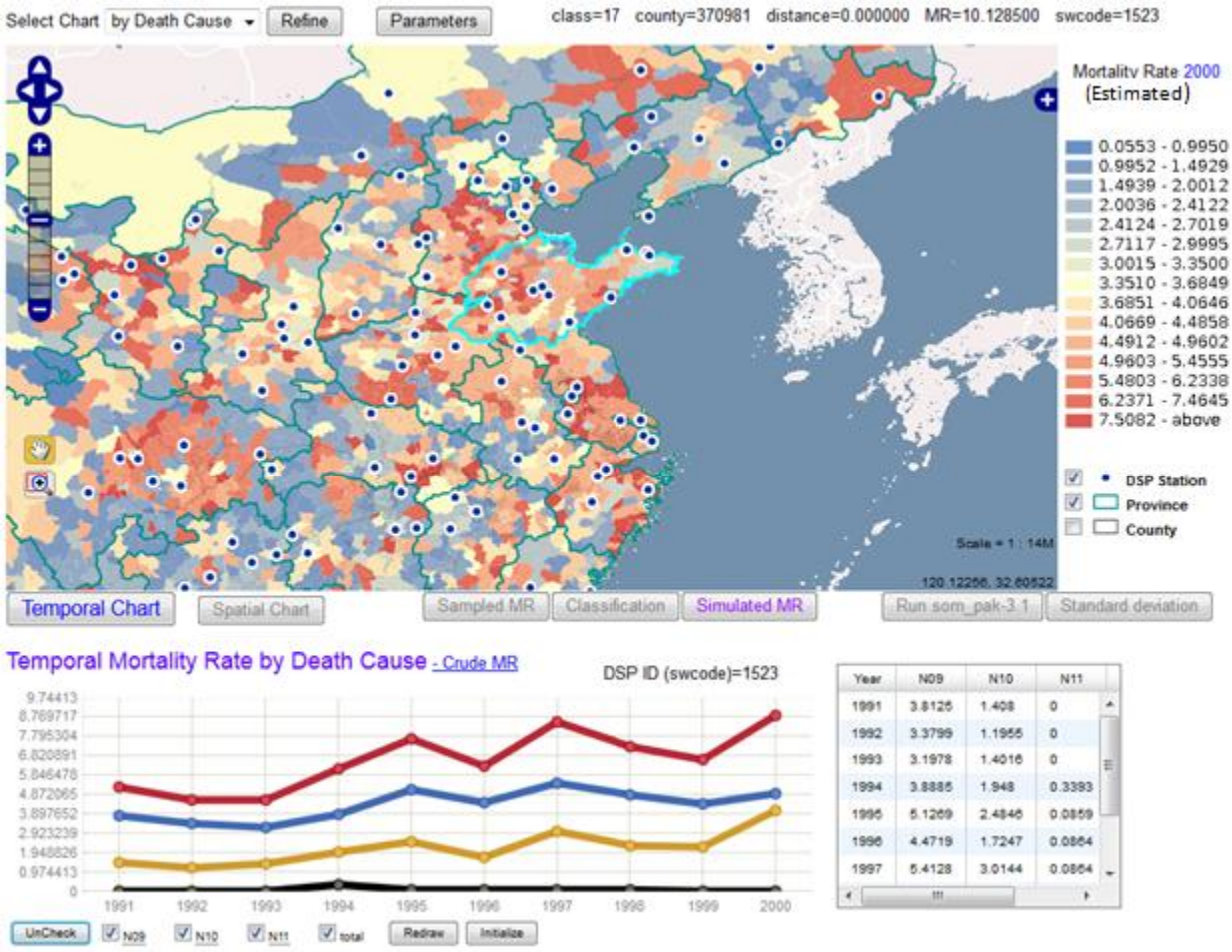


Figure 13

Users examine the temporal changes of the mortality rates of orange or red colored areas—i.e. areas with the high potential rates of mortality.

CHAPTER 4

CASE STUDY

The case study is designed as a CyberGIS application in which spatiotemporal characteristics of mortality rates in China (mainland) are examined for the period of 1991–2000. This chapter addresses spatial analysis needs of a set of China mortality data, and describes specific analysis methods developed, and corresponding findings. A specific focus is placed on the evaluation of spatial data analysis methods and a CyberGIS user interface that supports interactive exploration of spatiotemporal data attributes.

4.1 DATA

There are two different datasets available for this study. First, the China national mortality dataset covers 1% population in the mainland of China consisting of 2420 counties. This dataset was collected in 145 counties by using multiple stratified random sampling for the period of 1991 to 2000 (Figure 14) (Yang *et al*. 2008).

In these 145 sampled counties, each station uniquely identified by its four-digit ID is allocated for recording information about every death. Each mortality record includes gender, age, occupation, ethnicity, education, marriage statues, death date, death place, and death causes. The mortality samples were chosen to represent regional population distributions, urban and rural areas, age and sex, and eastern, middle, and western regions (Yang *et al*. 2008). The second dataset is China 1990 population census collected at the county level. Each county includes 40 socioeconomic and demographic data attributes representing population, age-sex structure, education, marital status, the number of total births and death, industrial/economy activity and occupation (Table1). A spatiotemporal database of the ten-year mortality data and

socioeconomic and demographic attributes has been established, and is integrated within the CyberGIS application.

Figure 14

Study area is the mainland of China. Sampled counties with mortality samples are represented in grey-filled counties.

Category	Attribute	Normalized by
• population	# of pop	area
	# of households	area
	# of non-agricultural households	total households
	# of agricultural households	total households
	# of immigrants since 1985	total pop
• sex	# of male pop	total pop
	# of female pop	total pop
• age	age 0 - 4	total pop
	age 5 - 14	total pop
	age 15 - 39	total pop
	age 40 - 64	total pop
	age 65 above	total pop
• education	# of pop having college degree	pop over age 15
	# of illiterate pop	pop over age 15
• marital status	Never Married	pop over age 15
	Married	pop over age 15
	Widow	pop over age 15
	Divorced	pop over age 15
• birth/death	# of births between 1989 - 1990	total pop
	# of deaths between 1989 - 1990	total pop
• industrial /economic activity	# of total employed pop	pop over age 15
	# of pop employed in agriculture	total employed pop
	# of pop employed in industry	total employed pop
	# of pop employed in mining, prospecting	total employed pop
	# of pop employed in construction	total employed pop
	# of pop employed transport, posts, telecommunications	total employed pop
	# of pop employed in commerce supply and marketing	total employed pop
	# of pop employed in real estate, utilities, residential services	total employed pop
	# of pop employed in medicine, health care, sports, welfare	total employed pop
	# of pop employed in education, culture, arts, radio, television	total employed pop
	# of pop employed in science, technology	total employed pop
	# of pop employed in finance, insurance	total employed pop
	# of pop employed in government, party, and NGOs	total employed pop
• occupation	# of professional and high-level technical personnel	total employed pop
	# of officials/managers in gov't, party, business, & NGOs	total employed pop
	# of clerical personnel	total employed pop
	# of employees in commercial sector	total employed pop
	# of employees in service sector	total employed pop
	# of workers in agriculture, forestry, husbandry, fisheries	total employed pop
	# of workers in manufacturing, construction, transport, etc.	total employed pop

Table.1

40 socicoeconomic and demographic attributes in China 1990 population census. These attributes are available in every county. Each attribute in the 2nd column is divided by another attribute of the 3rd column in the same row.

4.2 CHALLENGES OF SPATIAL DATA ANALYSIS

The epidemiological transition from infectious disease and perinatal conditions to chronic diseases and injuries has occurred at a much faster pace in China than in many western countries (Yang *et al*. 2008). Therefore, it is necessary to identify surveillance areas where populations are excessively exposed to certain disease risks. Identifying areas of high mortality rates would help public health officials to make informed decisions on where to allocate medical resources. To perform such a task, however, only available datasets are the mortality and census data (see section 4.1 for details). The primary question that this study addresses is how to analyze these datasets to represent the mortality rates of non-sampled counties and identify high risk areas of certain diseases in the mainland of China.

4.3 METHODOLOGICAL APPROACH

Given the high dimensionality and complexity of the available data, our approach is based on exploratory data analysis (EDA). The principle of EDA is to let data speak for themselves by imposing no *a priori* hypothesis (Gould 1981). By utilizing statistical tools and information visualization such as cartographic maps, tables, histograms, scatter plots, and charts (Harris 1999), EDA employs visual abstractions to reveal "potentially explicable patterns" of data (Good, 1983). The general purpose of EDA is to represent data in an ordered fashion such as based on clustering structures and relations among data elements (Kaski and Kohonen 1996). Several EDA methods focus on user interactivity through data visualization. They allow users to explore various data views in response to their parameter selections, which are often enabled by computation for data processing (see Cleveland, 1993; Buja *et al*., 1996; Guo 2005). The findings based on EDA can guide users to suggest explanations, create formal hypothesis and

theoretical constructs, and present data in a form that is easily understandable (Messner *et al*. 1999; Guo 2005).

Kaski and Kohonen (1996) illustrated a case in which an exploratory data analysis can be used instead of constructing a prediction model. Data and knowledge required for formulating accurate prediction models could often be too costly to acquire. Even though a model can be developed based on limited data and related knowledge, the model may fail to achieve desired quality. In this particular situation, EDA is also appropriate. Although predicting mortality rates of non-sampled counties is desirable in this research context, it would not be possible to construct a reliable model given our limited data. Therefore, instead of developing a model for prediction, we use EDA to produce useful hints on identifying areas of high mortality rates.

Our exploratory spatial data analysis encompasses two interrelated methods. The first is the self-organizing map (SOM) that focuses on grouping socioeconomically and demographically similar counties and visualizing the information of the mortality data in each group through statistical summary. The second is a straightforward spatial classification (SSC) method that aims to reveal the entire patterns of the mortality rates of the mainland of China based on the similarity of socioeconomic and demographic factors. To measure the similarity between high dimensional socioeconomic and demographic attributes of counties, both methods use Euclidean distance (see section 4.3.1 for details).

4.3.1 Measuring the similarity of socioeconomic and demographic attributes

Data Preprocessing. In measuring the similarity between high dimensional socioeconomic and demographic attributes of counties, data preprocessing plays an important role. The first data-preprocessing step is that each attribute is divided by another attribute as

indicated in Table 1 (see Dailey 2006). The next step is to use the minimum and maximum values for normalizing all of the attributes. The formula is given below:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

This normalization method scales all of the attributes to the range [0,1] for a fair comparison between them (see Guo 2005; Vesanto *et al*. 1999).

A Similarity Measure. To measure the similarity or distance between high dimensional socioeconomic and demographic attributes, *Euclidean distance* is used (see Kohonen 2001): for county x with n attributes - $x = (x_1, x_2, ..., x_n)$ and county y with n attributes $y = (y_1, y_2, ..., y_n)$,

$$d(x, y) = \| x - y \| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}.$$

The shorter the distance among the socioeconomic and demographic attributes of counties is, the more socioeconomically and demographically similar the counties are.

4.3.2 Similarity comparison based on EDA

Revealing the patterns of mortality rates of the entire mainland of China can be achieved by associating the mortality rates of the sampled counties to the non-sampled counties that are socioeconomically and demographically similar to the sampled counties. For example, Figure 15 illustrates how this process works.

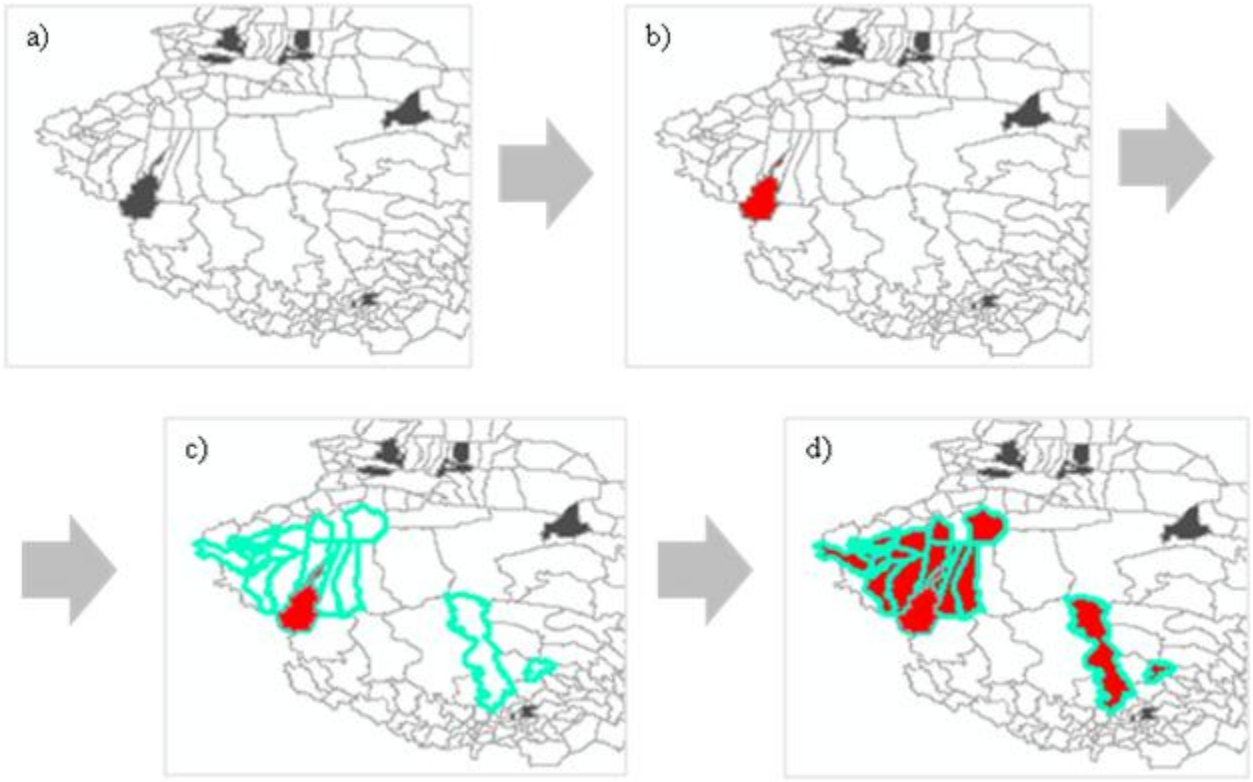


Figure 15

The mortality rate of a sampled county could be similar to the mortality rates of some non-samped counties that are socioeconomically and demographically similar to the sampled couty. The mortaltiy rates of the sampled counties can be representative of all the counties within the areas that have socioeconomically and dmeogrpahically similar population characteristics (outlined with bright blue).

There are sparsely sampled counties indicated by grey-color polygons (Figure 15a). Given a sampled county (red colored) that has a high mortality rate (Figure 15b), there could be non-sampled countries whose socioeconomic and demographic attributes are similar to those of the sampled county. The similar counties are highlighted using bright blue color (Figure 15c). Geographic neighbors may be similar in such attributes, which can be explained by the first law of geography: everything is related to everything else, but near things are more related than distant things.

Once socioeconomically and demographically similar counties are identified, a mortality rate of a sampled county could be similar to mortality rates of non-sampled counties within the similar counties. Figure 15d provides an example that a set of non-sampled counties are identified with high mortality rates because these counties are socioeconomically and demographically similar to the sampled county that has a high mortality rate. In other words, the high mortality rate of the sampled county is chosen as a representative mortality rate of all of the similar counties (outlined with bright blue). Without losing generality, we hypothesize that the mortality rate of a sampled county can be associated to non-sampled counties that are socioeconomically and demographically similar to the sampled county. Because the research focus of this thesis is placed on spatial analysis methods and CyberGIS, it is appropriate to use 1990 census data (see Table1) to assess the socioeconomic and demographic similarity although socioeconomic and demographic status may change during the 10-year span (1991 – 2000).

4.3.3 EDA for knowledge discovery using self-organizing map

Self-organizing map (SOM), also known as Kohonen network, is a type of artificial neural networks (ANNs) and involves unsupervised learning (see Skupin and Agarwal 2008). It projects high dimensional data onto a low dimensional space (usually 2D space) through the use of self-organizing networks while preserving nonlinear relations of data. Kohonen (2001) provided detailed mathematical descriptions of SOM.

A typical SOM is composed of input vectors $X_i = [X_{i1}, X_{i2}, ..., X_{in}]$, where i denotes the number of vectors and n denotes the number of attributes (see Figure 3), and a two-dimensional array of nodes in output space (SOM) (Figure 16).

attribute (n)

input vector (i)

	var1	var2	var3	var4
geographical unit 1	x_{11}	x_{12}	·····	x_{1n}
geographical unit 2	x_{21}	x_{22}	·····	x_{2n}
⋮	⋮	⋮		⋮
geographical unit i	x_{i1}	x_{i2}	·····	x_{in}

Figure 16

When *n* is large, data are high dimesnional.

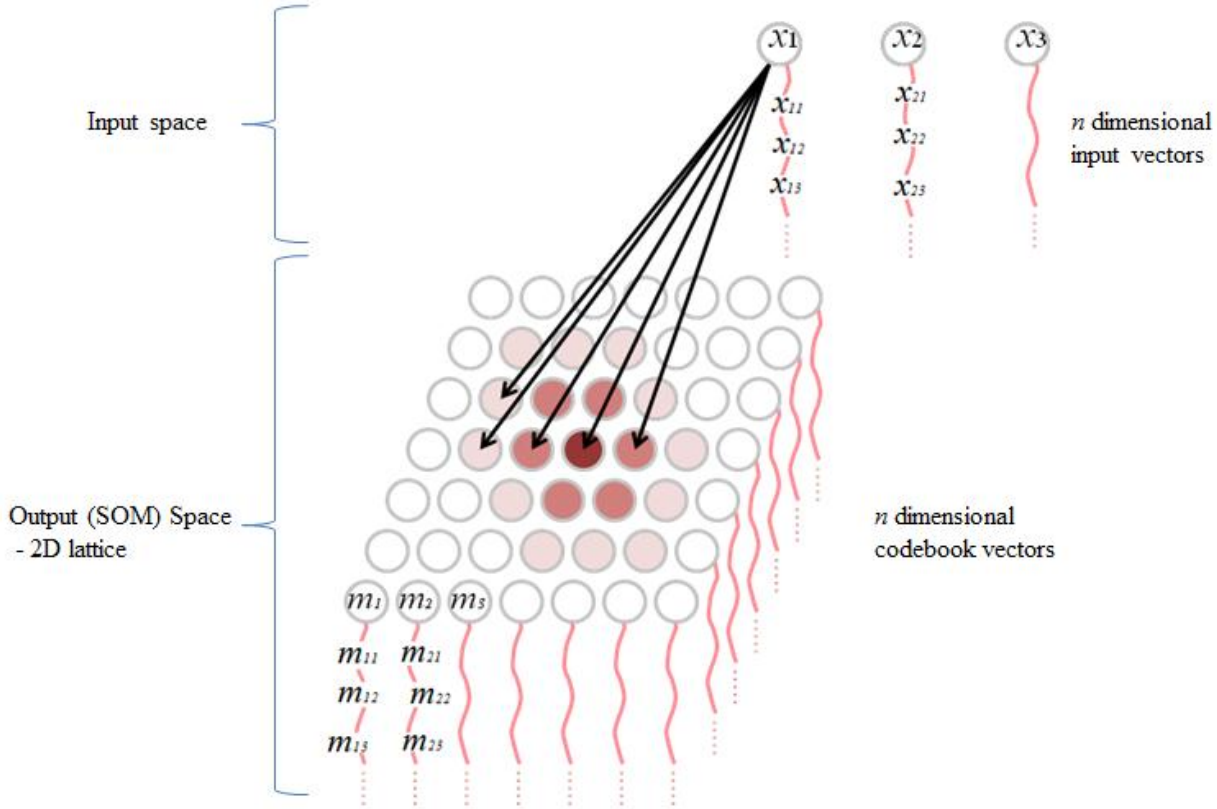


Figure 17 A structure of self-organizing map

The lattice type of the array is usually rectangular or hexagonal. With every node *i*, *n*-dimensional vector, called a codebook vector $m_i = [m_{i1}, m_{i2} ..., m_{in}]$ is associated, where *n* is equal to the number of attributes of input vectors (Figure 17). In other words, the dimensionality of each node's codebook vector m_i is identical to that of the input vectors x_i. Each input vector is connected to each node's codebook vector in parallel.

At the initialization step of the SOM algorithm, codebook vectors of each node are typically assigned with random numbers. Then, an iterative learning process is executed. At the first step, one vector among input vectors x_i is selected randomly, and then the chosen input vector is compared with all nodes' codebook vectors to find its most similar node. To measure the similarity between the chosen input vector and the codebook vector of each node, *Euclidean Distance* $|| x - m_i ||$ is calculated. At the second step, the node closest to the chosen input vector is made to define the best matching unit (BMU) that is also the most similar node to the chosen input vector. Once BMU is found, the chosen input vector is assigned to the BMU and its neighbors are adjusted to make them more similar to the chosen input vector. At the third step, the closer a node is to the BMU, the more its codebook vectors get-altered to become more similar to the chosen input vector. The process of becoming adjusted to be more similar to the chosen input vectors is often referred to as a learning process. These three steps are repeated for a large number of times. As a "rule of thumb," Kohonen suggested that the number of repetition should be more than 500 times the number of nodes. We demonstrate how the SOM algorithm works with the socioeconomic and demographic data in China as follows.

Through the learning process, SOM classifies counties that are represented based on the 40-dimensional socioeconomic and demographic attributes and visualizes the classification of counties on 2D (SOM) space. Commonly used hexagonal lattice type is adopted. To choose the dimension of SOM is often based on experimental processes (e.g. Oyana *et al*. 2005). When the size of nodes is too small, too many sampled counties are assigned into a few nodes. On the other hand, if the size of nodes is too large, it creates too many empty nodes. Therefore, in this thesis, a mid-range dimension of 7 by 7 nodes is configured, which leads to meaningful findings (see Figure 22 for details). For the calculation of SOM, above all, all of the attributes are preprocessed (see section 4.3.1 for details). For each iteration, one of the counties is chosen randomly for the learning process, and this county is assigned to a node if the node's codebook vector is the closest (most similar) to the county's attributes. Multiple counties can be assigned to one node while some nodes may be empty.

After the iterative learning process is done, counties having similar socioeconomic and demographic attributes are mapped to be close to one another in the SOM space. In other words, in each of the 49 classes, socioeconomically and demographically similar counties are grouped together and counties belong to nearby nodes are similar to each other. The terms "node," "class," and "type" are interchangeably used in this thesis. Consequently, 49 different groups are created. Each of the counties in the mainland of China is represented by one of the 49 population groups. Therefore, the 49 different population groups cover all population characteristics of China. Figure 5 is captured from the CyberGIS application (see section 3.2 for a complete description).

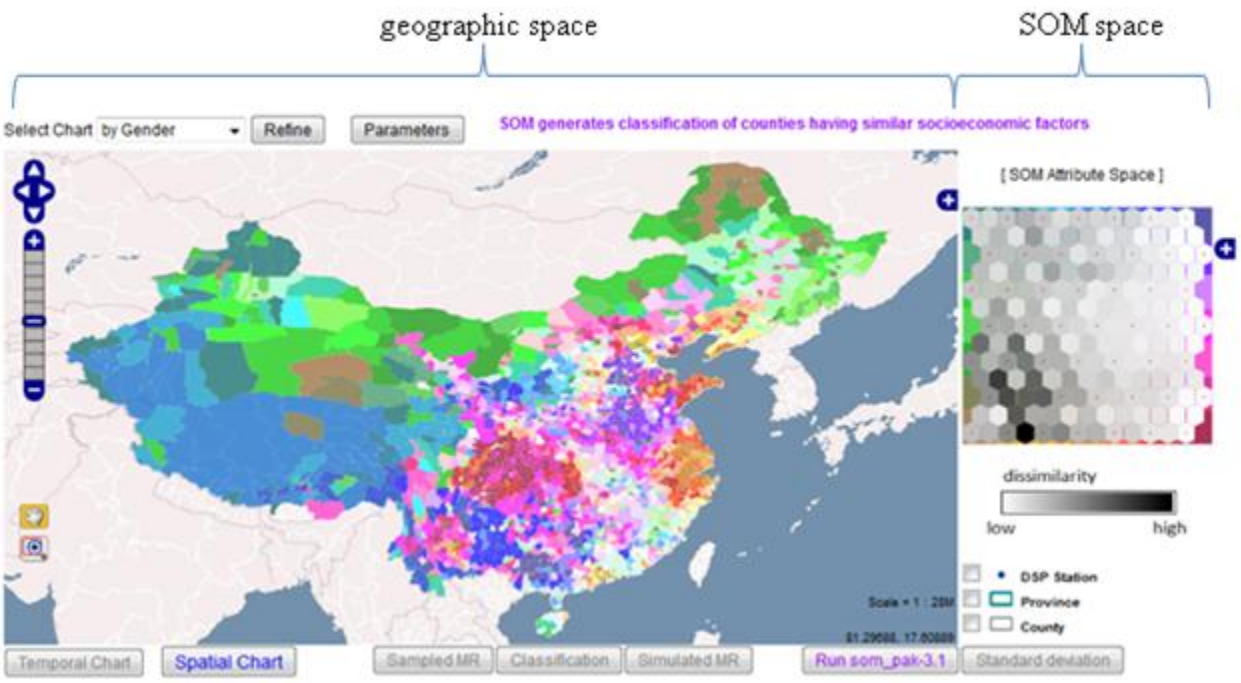


Figure 18

Geographic space on the left and the SOM space on the right. Counties in the geographic space are represented by the same color as the node to which the counties belong. The SOM space is comprised of two layers (see Figure 19).

An image on the right is the 7 by 7 SOM space. The SOM space specifies what counties belong to each class, but it does not depict geographic relationships among counties. To represent a spatial distribution of the 49 groups, the geographic space is linked to the SOM space, and data visualization and interactive mapping help connect between those two spaces.

In the SOM space, two layers are overlapped (Figure 19). A bottom layer is a color representation (Figure 19a) (Kohonen, 2001), and an upper layer is a U matrix representation (Figure 19b).

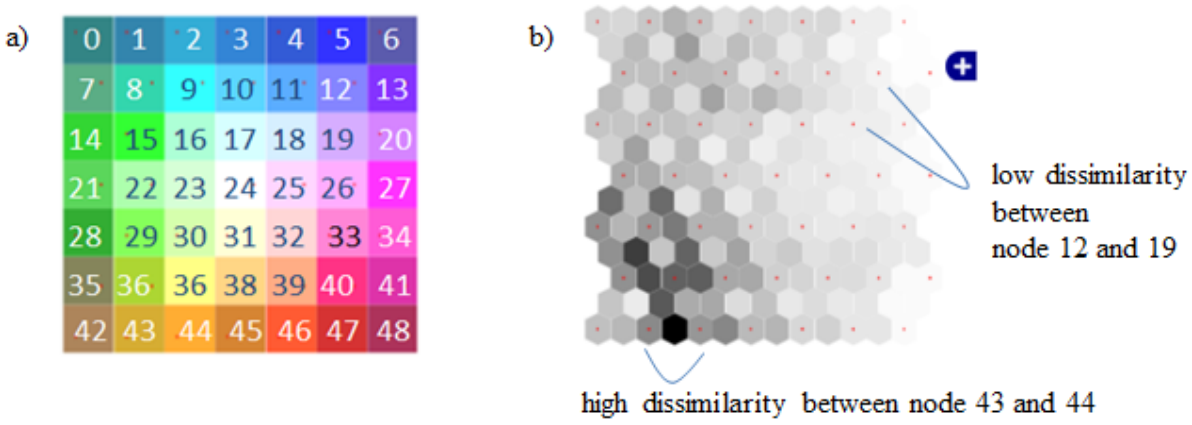


Figure 19

a) Color representation of the SOM space: neighboring nodes are represented by similar colors. Counties belonging to neighboring nodes are generally similar to one another. b) U-matrix represents the dissimilarity of counties belonging to neighboring nodes: a center of each node is represented by a red dot.

In general, counties that belong to the closely located nodes are socioeconomically and demographically similar to one another, whereas counties that belong to nodes located far away are socioeconomically and demographically dissimilar to one another. Therefore, neighboring nodes are visualized in similar colors—e.g. colors of nodes 40, 41, 47 and 48 are similar to one another. In contrast, nodes located far away are represented in more different colors—e.g. the red color of node 0 is the complementary color of the green of node 48.

In the geographic space, counties are also represented using the same color assigned to a node where the counties belong to in the SOM space (see the geographic space in Figure 18 and Figure 19a). For example, counties belonging to node 48 colored by red in the SOM space are also represented by red colored counties in the geographic space. Therefore, counties with similar colors in the geographic space mean that they are similar to one another in terms of socioeconomic and demographic variables.

U matrix focuses on representing the dissimilarity among neighboring nodes. A darker grey indicates that there is significant socioeconomic and demographic dissimilarity among neighboring nodes. Conversely, a lighter grey represents less dissimilarity among the neighboring nodes. For example, nodes 43 and 44 are neighboring to each other in the SOM space, but there is a substantially high dissimilarity between the counties that belong to these two nodes. In contrast, the dissimilarity between the two neighboring nodes 12 and 19 is relatively low.

In the CyberGIS application, the connection between the SOM and geographic spaces can be interactively examined. When a user clicks one of the SOM nodes, counties belonging to the clicked node are highlighted in the geographic space. For example, as a user clicks node 48 colored by red on the SOM space (see Figure 19a), counties belonging to the node 48 are highlighted in the geographic space (Figure 20).

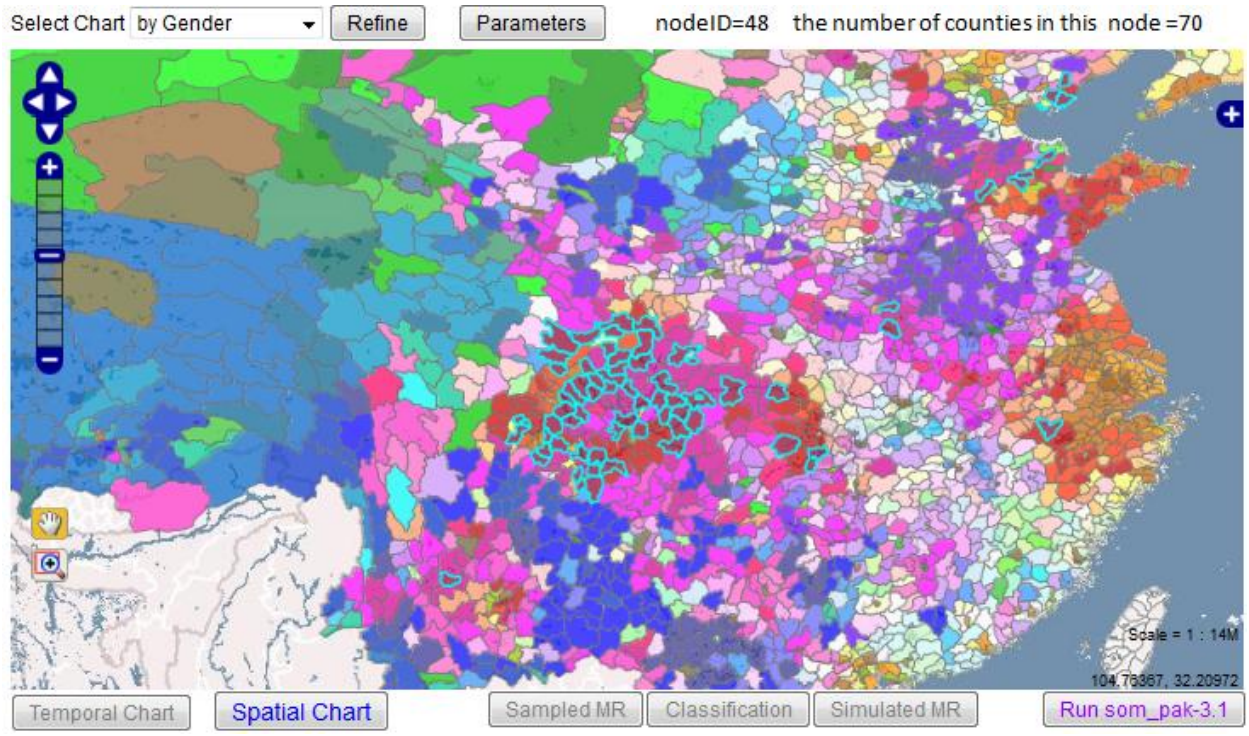


Figure 20 Linkage between the SOM and the geographic spaces

The user interface also shows the number of counties belonging to the node 48. In the same manner, the degree of the dissimilarity among counties in the geographic space can be also examined by highlighting counties in response to clicking each node in the U matrix (see Figure 19b).

In the geographic space, the classification and clustering of socioeconomically and demographically similar counties are generated by a SOM process (Figure 21).

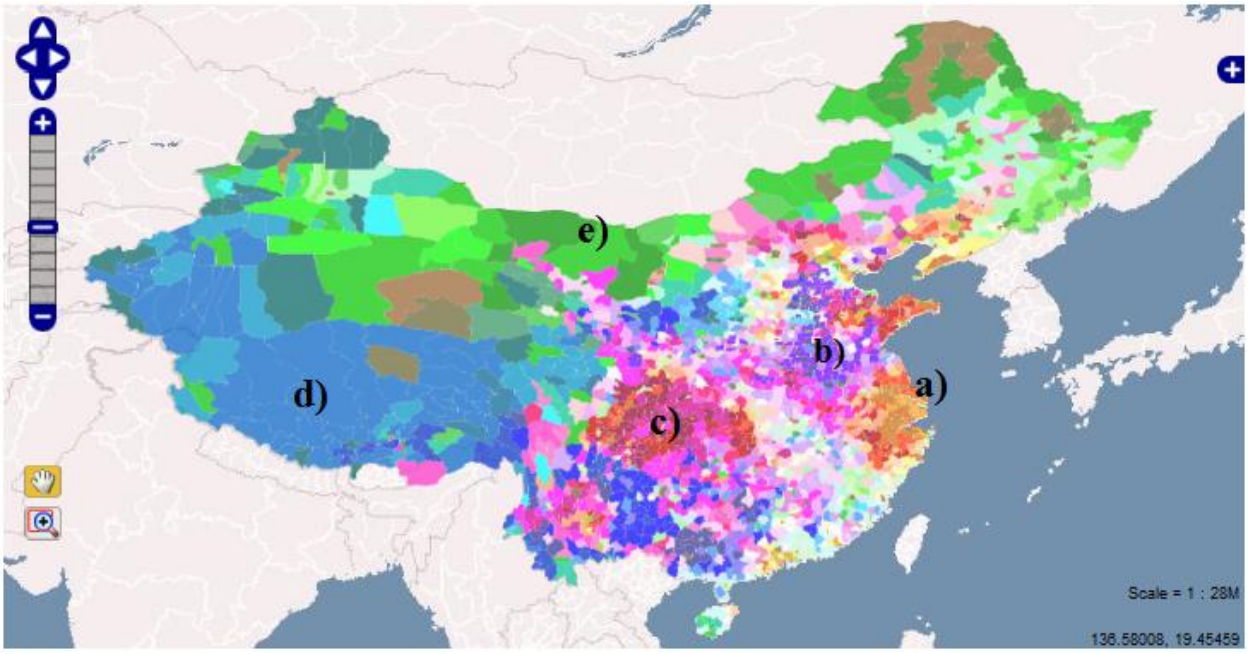


Figure 21

Major clusters of socioeconomically and demographically similar counties.

a) cluster: orange colored counties in the eastern coastal areas
b) cluster: purple and light pink colored counties in inner-land areas of east coast
c) cluster: red and pink colored counties in the center region.
d) cluster: blue colored counties in southwest areas.
e) cluster: green colored counties across northwest areas to northeast areas

Clusters of *a*, *b* and *c* are groups of similar counties in highly populated areas in the southeast, central, and southwest parts of China. On the other hand, *d* cluster contains similar counties where mountainous and high plateaus are located. *e* clusters seems to represent a transition group

between *d* and other clusters. Therefore, these clusters show that counties having similar population characteristics have geographic dependencies. Based on the classification using SOM, information about a distribution of mortality rates of each year is added to the application (Figure 22).

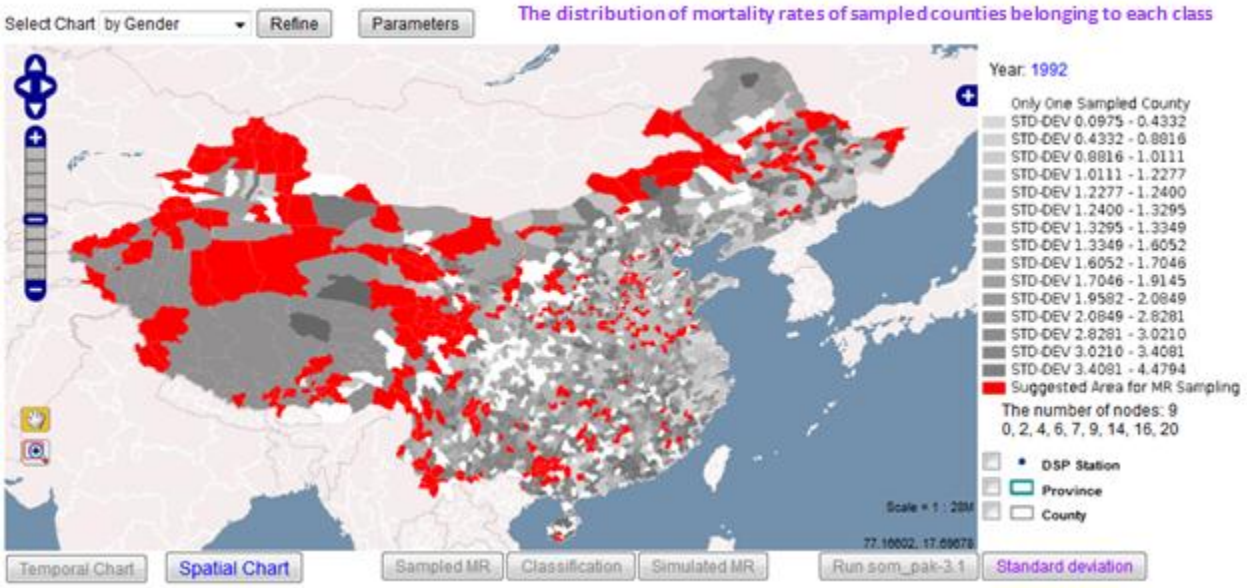


Figure 22

Counties are shaded to represent the difference of mortality rates belonging to each class. Darker counties indicate that there is a significant difference among the mortality rates of sampled counties belonging to the same class. Lighter color represents that the mortality rates of sampled counties belonging to the same class do not have any significant difference. White colored counties are the members of classes having only one sampled county. Red colored counties belong to classes where no sampled county is included. Different views of this map are provided in Figures 24, 25, 26 and 27.

The image is generated in response to a user's selection of input parameters (see section 3.2.1 for details). Figure 23 shows the selected input of mortality data—i.e. the total mortality rate and the year 1992. The 145 sampled counties were selected to represent different types of population characteristics of the mainland of China (see section 4.1).

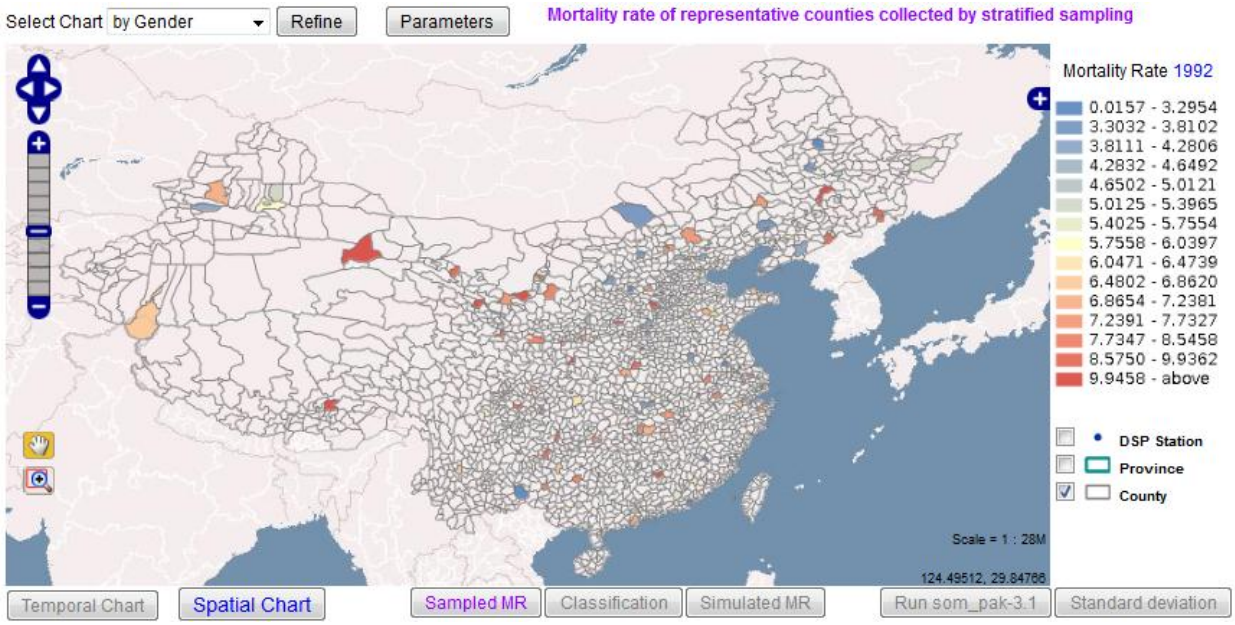


Figure 23

It represents the total mortality rates of sampled counties in 1992. They are inputs for the map shown in Figure 22.

Our EDA based on SOM sheds light on how comprehensively these samples can be representative of different population characteristics in relation to socioeconomic and demographic factors. Red colored counties on the map belong to classes where population characteristics do not have any matching sampled mortality data. In the legend area of Figure 24, nine classes where red colored counties belong to are displayed. Based on the 49 classes generated by SOM (see Figure 18), each of the nine classes is socioeconomically and demographically different from one another, whereas counties belonging to each same class are comparatively similar to one another. A group of counties in each of the nine classes respectively reflects distinctive population characteristics, but none of the counties belonging to those classes are sampled. For example, class 0 has 24 socioeconomically and demographically similar counties that are different from counties belonging to other classes (Figure 24). None of the 24

counties, however, has any representative mortality rate sampled. Therefore, it would be suggested that at least one of the counties belong to the class 0 should be sampled to have a better coverage reflecting all different population characteristics. Through the classification based on SOM, one or more counties in each class of counties suggested in Figure 24 are candidates to be sampled to have a more comprehensive coverage of various population characteristics.

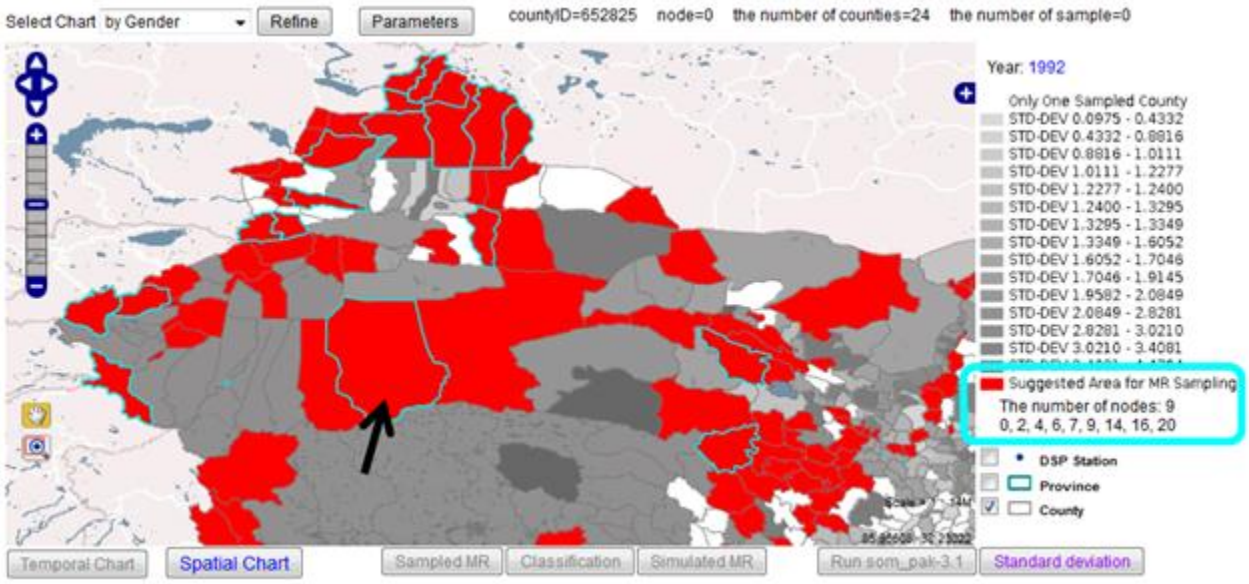


Figure 24

There are 9 classes where counties do not include any representative mortality rate. Among counties belonging to classes 0, 2, 4, 6, 7, 9, 14, 16 and 20, one or more counties in each of these classes are suggested to be sampled. The suggested sites for sampling of mortality data can be examined interactively by users. For instance, the class 0 is clicked above (indicated by an arrow). Once it is clicked, all 24 counties belonging to the class 0 is highlighted (see the info on the top of a map). Among them, one or more counties may be sampled.

A cost-effective sampling strategy may also be suggested, by taking only one sample from each of the 49 classes since the 49 population classes could be regarded as the representatives of various population characteristics in China. Counties belong to the class

having only one sampled county is represented in white on the map. For example, Figure 25 shows that class 48 has 70 similar counties, and among them the mortality rate of 7.233 of the clicked county would be a representative of all of the highlighted counties. Therefore, white-colored counties would likely be cost-effective sampling areas.

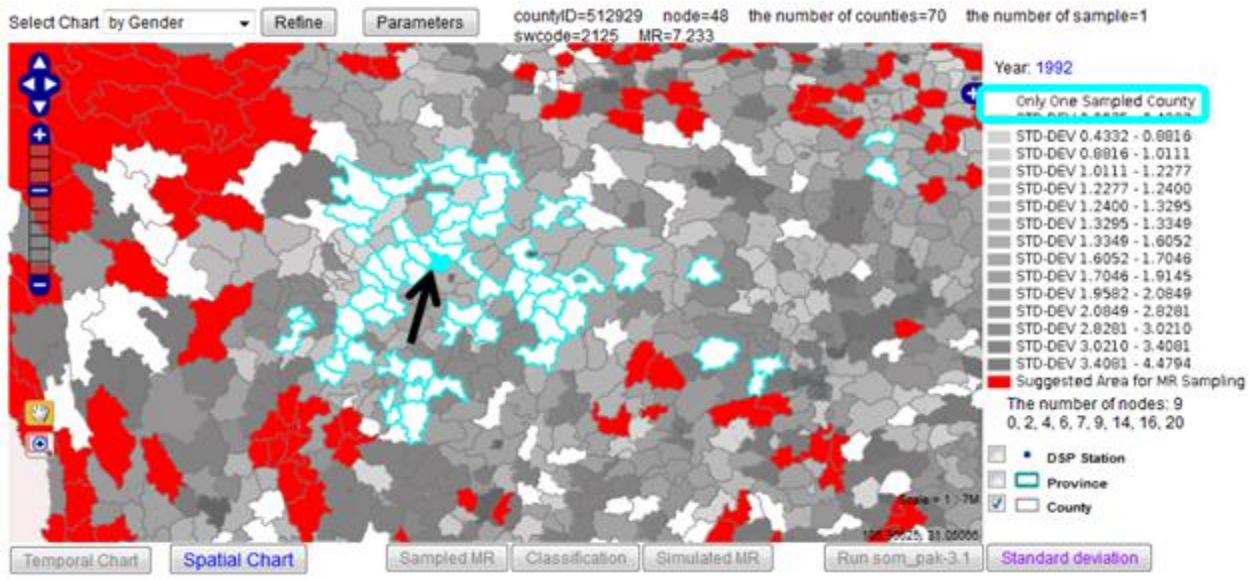


Figure 25

When the user clicks one of counties on the map, all sampled counties belonging to the same class having the clicked county are filled in bright blue. At the same time, non-sampled counties belonging to the same class are outlined with bright green. In addition, information about the clicked county is displayed. A clicked county is highlighted using an arrow. A county (ID: 512929) is clicked (as shown in Figure 19a). Class 48 includes 70 counties, and among them only one is a sampled county. In this case, the sampled county is clicked, so it also displays the station code (swcode: 2125) of the sampled location and mortality rate, 7.233 of the clicked county.

Investigating the difference of mortality rates of counties belonging to each SOM class provides insights to understand the relationships between mortality rates, and socioeconomic and demographic factors in a comprehensive way. For each class having more than two sampled counties, counties are shaded to represent the difference among mortality rates of sampled

counties belonging to the class. In this case, the standard deviation of mortality rates of sampled counties is calculated. The darker areas represent the larger standard deviation in the same class. For example, Figure 26 shows that there are 52 counties in class 45, and among them 5 counties are sampled for death records. When the range of mortality rates in 1992 is considered in Figure 23, the difference of mortality rates of the 5 sampled counties is relatively small with the standard deviation - 0.8816 - of the 5 sampled counties, which suggests a strong correlation between mortality rates, and socioeconomic and demographic factors in the 52 counties belonging to the class 45. In other words, for these counties, mortality rates tend to be similar in the areas that are socioeconomically and demographic similar to each other.

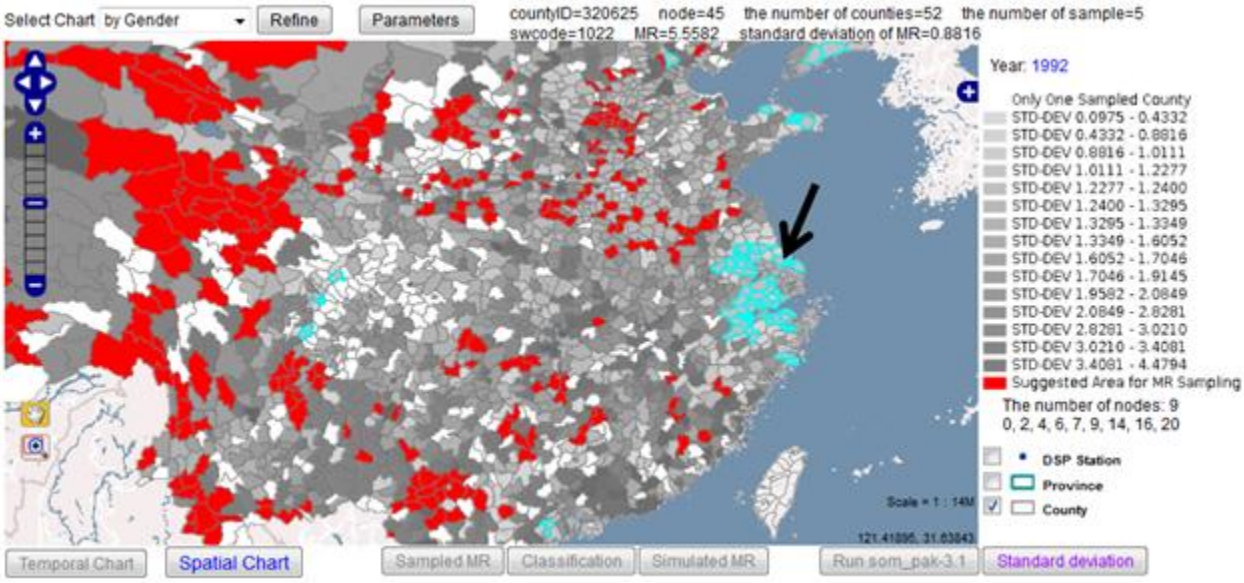


Figure 26

When the standard deviation is small, it means that there is a strong correlation between mortality rates, and socioeconomic and demographic factors. Once county (ID: 320625) is clicked (indicated by an arrow), all the highlighted counties are members of class 45. Class 45 has 52 counties as its members. Among them 5 counties were sampled. Standard deviation of mortality rates of all sampled counties is 0.8816 (see top of the map)

In contrast, some classes are revealed to have large standard deviations of mortality rates of sampled counties. For example, Figure 27 shows that there are 82 counties in class 5. Among them, 4 counties were sampled for mortality records, and the standard deviation—3.3616—is relatively large. In this case, it is hard to tell that mortality rates are likely to be similar within the counties that are socioeconomically and demographically similar to one another. In the counties belonging to the classes where significant difference of mortality rates of their sampled counties is observed, other factors than the socioeconomic and demographic ones may play a more significant role. For example, within counties belonging to the same class, there could be some counties that have land polluted by industry, leading to higher mortality rates than those of counties without land pollution.

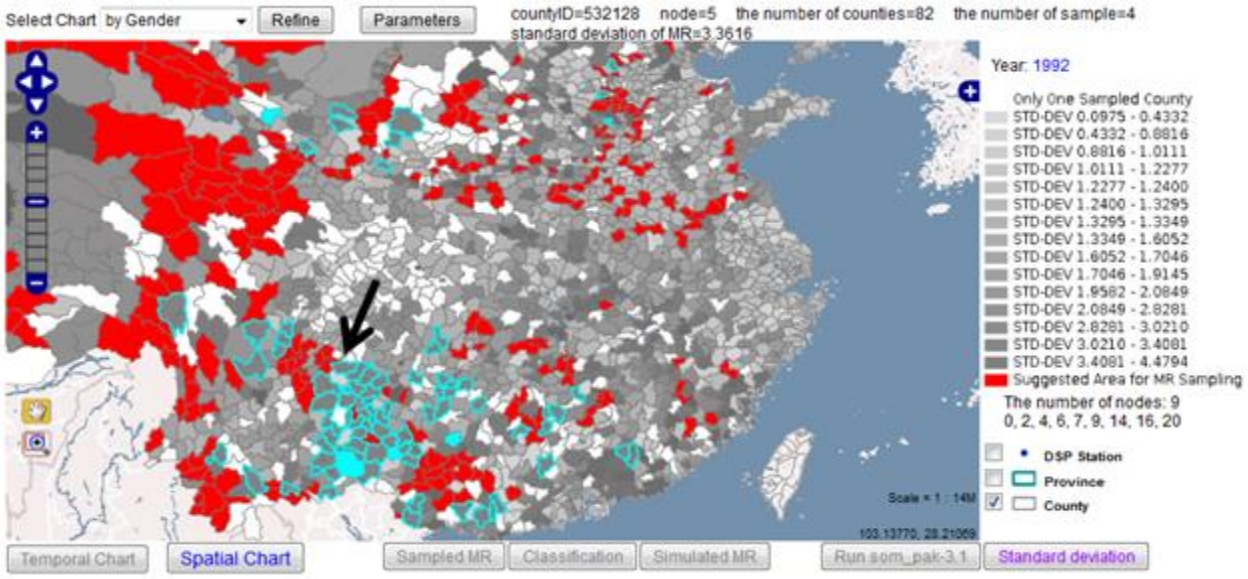


Figure 27

When standard deviation is large, it means that there may be other factors affecting mortality rates in addition to socioeconomic and demographic factors. County (ID: 53128) is clicked (indicated by an arrow). All the highlighted counties are members of class 5. Class 5 has 82 counties. Among them 4 counties were sampled. Standard deviation of the mortality rates of all sampled counties is 3.3616 (see top of the map).

4.3.4 The concept of straightforward spatial classification (SSC).

Based on the principles of EDA, the patterns of mortality rates of the entire mainland of China can be extrapolated from the mortality rates of the sampled counties. It would be desirable to associate a mortality rate of each sampled county with the mortality rates of non-sampled counties similar to the sampled counties, and see if any particular patterns of the mortality rates of certain diseases can be revealed. Exploring spatiotemporal patterns of mortality rates in this study is designed as a two-part process. The first part is to classify all of the counties into classes, within each of which its population characteristics are similar. The second part is to associate, within the same class defined in the first part, a mortality rate of a sampled county to the mortality rates of non-sampled counties. An idealistic case for the association is where there is only one sampled county in one class. For example, Figure 28 shows only one sampled county that belongs to one class where counties are similar to one another. In this case, the mortality rate of a sampled county can be considered as a representative mortality rate of all highlighted counties that belong to the same class.

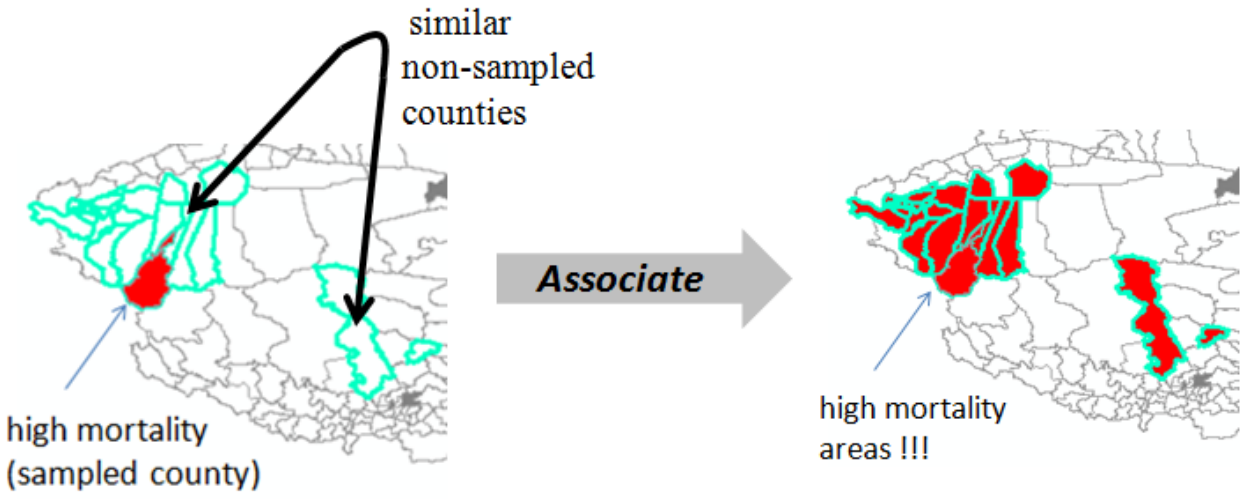


Figure 28

An idealistic case for associating a mortality rate of a sampled county with the mortality rates of non-sampled counties within the same class. A mortality rate of a sampled county would be considered as the representative mortality rate of all highlighted counties that belong to the same class. Counties belonging to the same class are outlined with bright green.

SOM is feasible for the classification of the high dimensional socioeconomic and demographic data, but is not suitable for the association. As described previously in section 4.3.3, the SOM method generates 49 different classes of population characteristics. However, given the SOM classification, there are some limitations on associating the mortality rates of the sampled counties to the mortality rates of the non-sampled counties. Specifically, SOM organizes the similar counties through an iterative learning process with multiple sampled counties often assigned to a same class. In some classes, none of the sampled counties are included. One sampled county per class is an idealistic case (Figure 28) for the association part, while the other two cases—no sampled county or multiple sampled counties in one class—requires further treatment. Suppose that a mortality rate is within a range between 2 and 8 across all counties (see the legend in Figure 29). As a result of SOM classification, there are, for instance, two sampled counties belonging to one class. One sampled county has the mortality rate of 7.3 and the other sampled county has the mortality rate of 7.4. These two mortality rates are likely to indicate relatively high mortality rates in the range between 2 and 8. Since the mortality rates of both of the sampled counties are relatively high (Figure 29.1a), the representative mortality rate of the counties belonging to the same class would be high (around 7.3 or 7.4) (Figure 29.1b). On the other hand, as a result of SOM classification, in one class very different mortality rates of sampled counties are assigned (Figure 29.2b)—e.g. one sampled county has a relatively high mortality rate (e.g. 7.2) while the other county has a relatively low mortality rate (e.g. 2.2). In this case, it is difficult to derive the representative mortality rate of the class.

In the SOM classification, there are even some classes without any sample county included. For example, Figure 29.3 shows two classes of counties. One class includes the sample 4.1 (outlined using pink color), and the other class including counties outlined using bright blue color. Within the former class, the sample, 4.1 can be associated as the mortality rates of all non-sampled counties. Therefore, the representative mortality rate of the counties outlined with pink color is 4.1. In contrast, in the latter class, there is no sample that can be representative of all non-sampled counties.

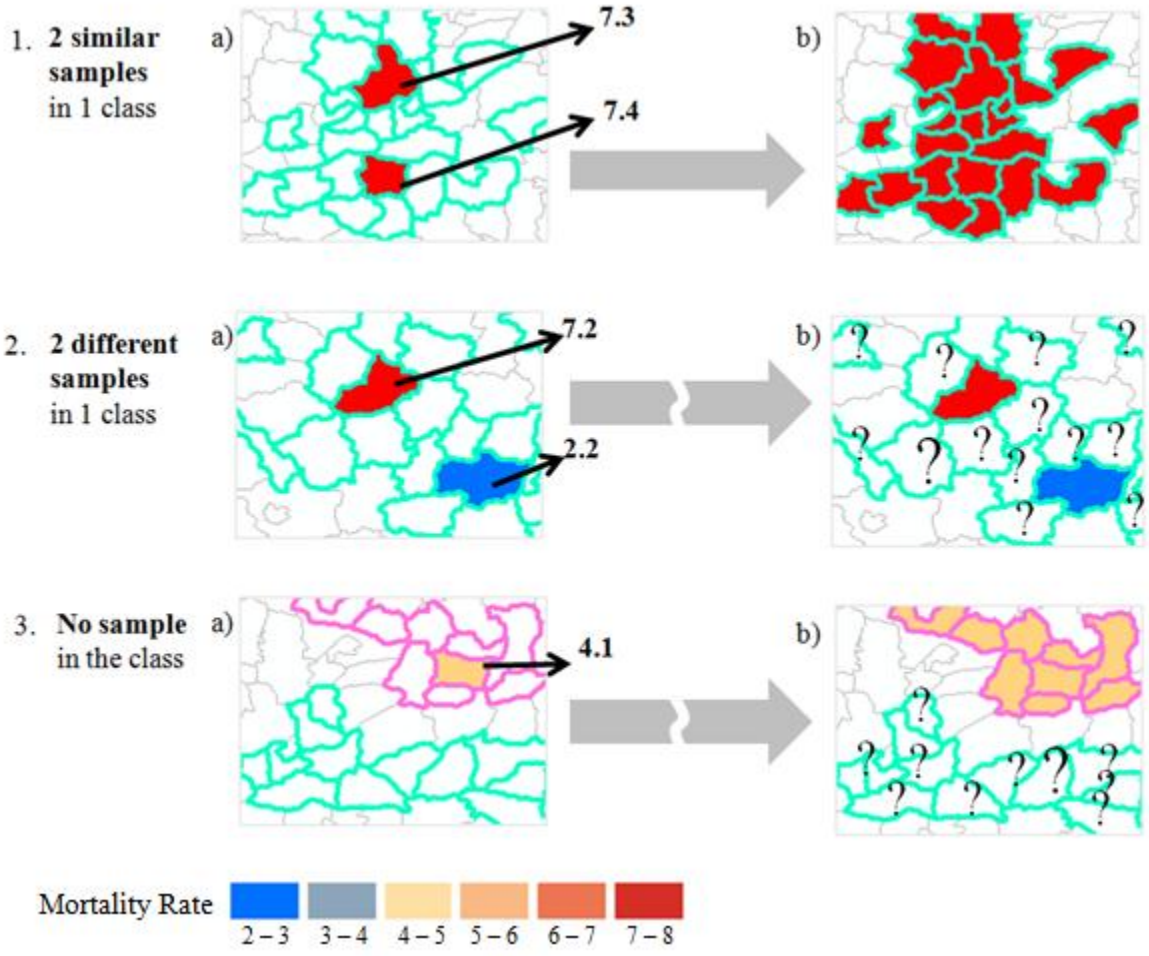


Figure 29

The classification generated by the SOM (left images) and the association based on the classification (right images): the association is not feasible in the second and third cases. Counties belonging to the same class are outlined using the same color. The socioeconomically

and demographically similar counties are organized together in each class. The numbers represent the values of mortality rates of sampled counties. In the first case, the representative mortality rate in the class would be reasonably between 7.3 and 7.4. In the second case, in the class, the two mortality rates of the two sampled counties are quite different. Therefore, it is not feasible to estimate the representative mortality rate of the class. In the third case, there are two separate classes. Among counties belonging to the class at the bottom, there is no mortality rate of the sampled county that can be representative of the class.

To address this association problem in the application of SOM and subsequently derive national distributional patterns of the estimated mortality rates, a method called straightforward spatial classification (SSC) has been developed. This method, like SOM, allows socioeconomically and demographically similar counties to be clustered together. Unlike the SOM, however, it is designed to always include only one sampled county in each class that may include multiple non-sampled counties.

The SSC method assures that each class includes one sampled county, which results into classes with the ideal association described in Figure28. The SSC measures the similarity between each non-sampled county and each sampled county. Then, for each non-sampled county, SSC searches each sampled county to find the most similar one. The shortest distance of 40 socioeconomic and demographic attributes between two counties means those counties are the most similar ones. As a result of this SSC process, while it is possible that a class includes only one sampled county and none of the non-sampled counties, it does not happen that any class includes only non-sampled counties or no sampled county based on the design of the SSC algorithm.

4.3.5 Description of SSC algorithm.

The algorithm developed for the straightforward spatial classification (SSC) method is described as follows. Inputs for this algorithm are 40 socioeconomic and demographic attributes of all of the sampled and non-sampled counties (Table1) and also the mortality rates of user-

defined diseases among user-specified population (see section 3.2.1 for details). Two groups are established as the basic data structure (Figure 30).

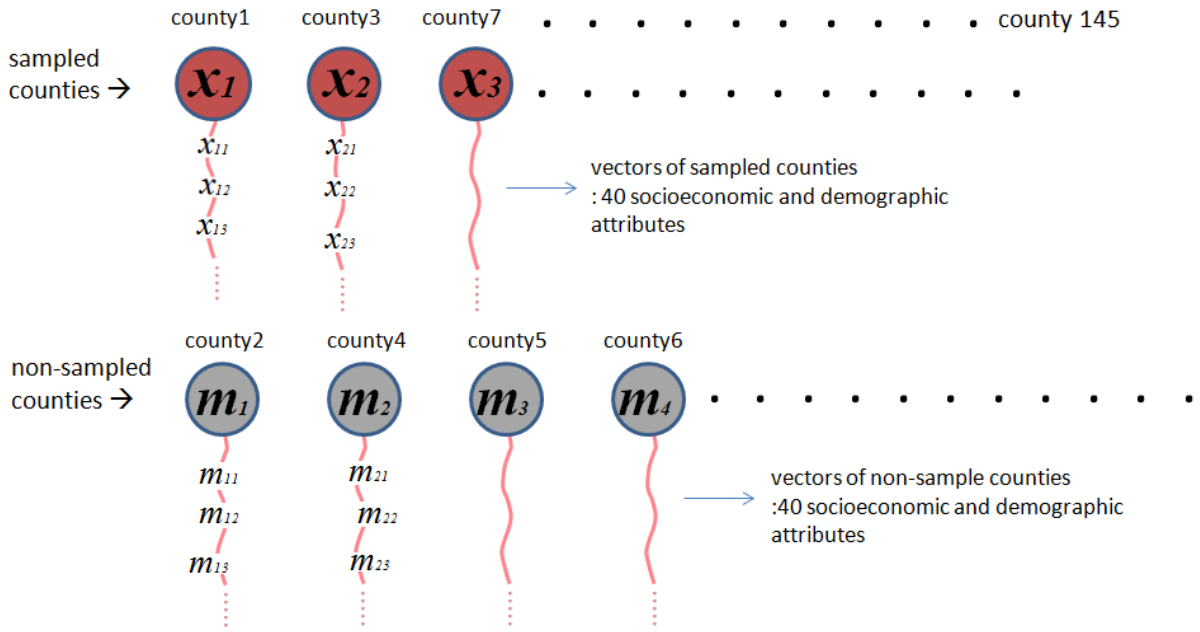


Figure 30 An illustrative diagram of straightforward spatial classification

The first group includes socioeconomic and demographic attributes of sampled counties x_i = [x_{i1} , x_{i2} , ..., x_{in}], where *i* is the index of sampled counties. The other group holds the same set of attributes of non-sampled counties m_i = [m_{i1} , m_{i2} ..., m_{in}],where *i* is the index of non-sampled counties. In both groups, *n* represents the number of socioeconomic and demographic attributes (*n*=40) (see Table1).

The algorithm includes multiple steps. At the first step, socioeconomic and demographic attributes of all of the counties are preprocessed as described in the section 4.3.1. The second step is to sort sampled counties to arrange the similar sampled counties next to each other (Figure 31).

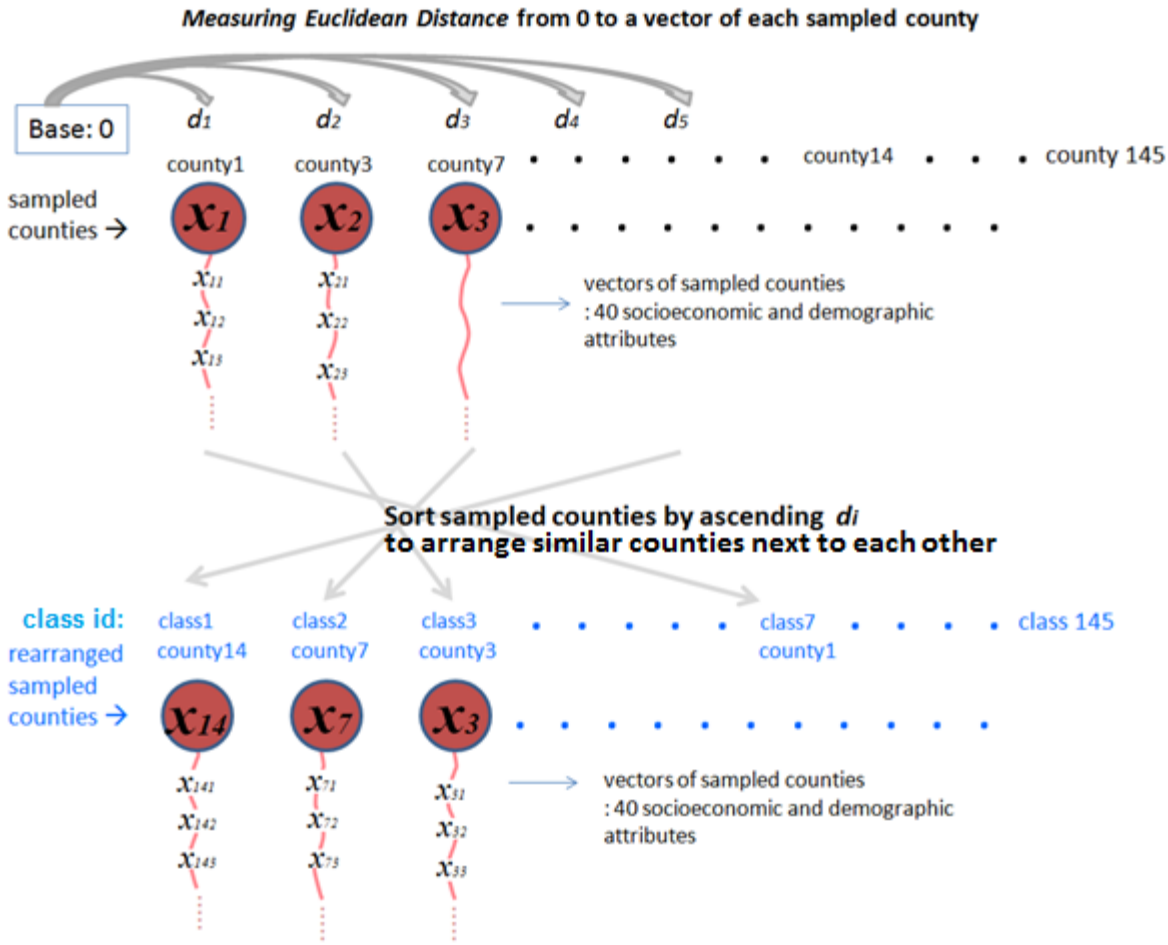


Figure 31 Sorting sampled counties by ascending d_i to place the similar sampled counties side by side. After the sorting is done, the class *id* is assigned to each sampled county.

To arrange the socioeconomically and demographically similar sampled counties side by side, *Euclidean distance* from 0 to a vector of each sampled county is calculated: for sampled counties having *n* socioeconomic and demographic attributes $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, where *i* is the index of sampled counties,

$$d_i = \| 0 - x_i \| = \sqrt{(0 - x_{i1})^2 + (0 - x_{i2})^2 + \ldots + (0 - x_{in})^2}.$$

Next, all sampled counties are sorted based on the ascending distance (d_i). After sorting is done, the class *id* is assigned to each sampled county in order from the smallest to largest distance. As a result, neighboring classes have similar sampled counties. For example, after the sorting, county7 of class2 is placed between county 14 of class1 and county3 of class3 since county7 is more similar to county14 or county3 rather than county1 of class7 (Figure31).

As the third step, the vector of every non-sampled county m is compared with the vector of all the sampled counties x_i to find one of the sampled counties that is most socioeconomically and demographically similar to each non-sampled county (Figure 32).

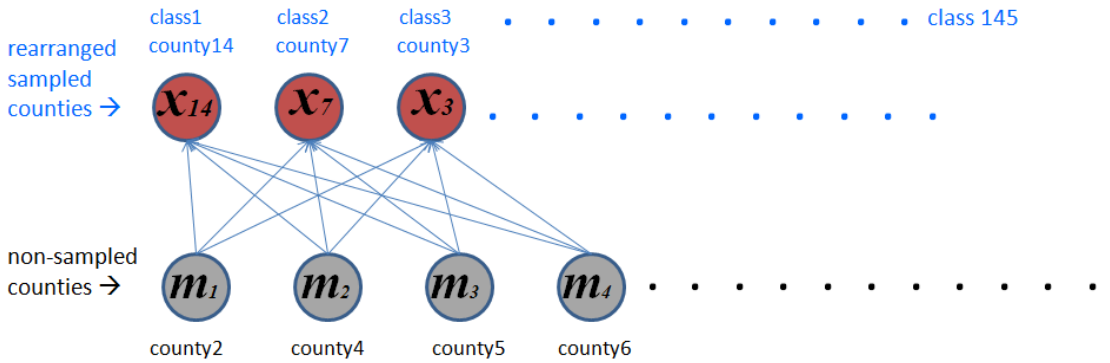


Figure 32 Measuring the similarity (distance) from the vector of every non-sampled county to the vectors of all of the sampled counties

To measure the socioeconomic and demographic similarity between sampled counties and non-sampled counties, the distance from the vector of every non-sampled county m to the vectors of all sampled counties x_i is calculated—i.e. *Euclidean Distance* $\| x_i - m \|$ is used in the same way as the SOM uses it for similarity measure. Then, for each non-sampled county the closest (the most similar) sampled county is chosen, and the chosen sampled county can be considered as BMU (Best Matching Unit) (Figure 33).

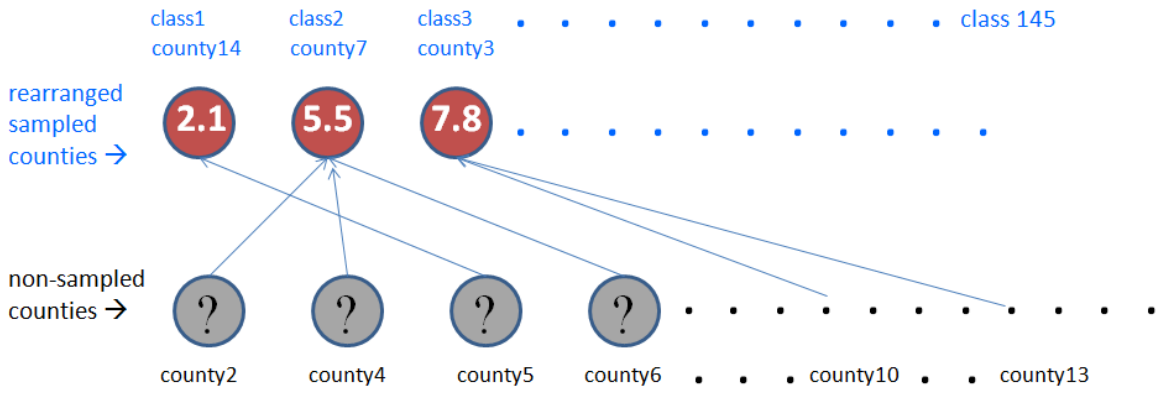


Figure 33

A value of each node represents a mortality rate. The BMU of county 5 is county14; BMU of county2, 4 and 6 is county7; BMU of county 10 and 13 is county3, etc. The values in red nodes represent the mortality rates of sampled counties. Question marks in grey nodes indicate that the mortality rates of non-sampled counties are unknown.

Once the BMU is identified, at the fourth step, the class *id* and a value of mortality rate of BMU are assigned to each non-sampled county. Figure 34 shows that the nodes of all of the sampled and non-sampled counties are rearranged according to the defined class *ids* and mortality rates. For example, non-sampled counties 2, 4 and 6 of class 2 hold the mortality rate, 5.5 of the sampled county 7 in the same class because the counties 2, 4 and 6 are most socioeconomically and demographically similar to the sampled county 7. In other words, the vector distance of the non-sampled counties 2, 4 and 6 is the closest to the sampled county7 among all the sampled counties. Consequently, the two parts of the classification and association are achieved: 1) classification: counties having similar socioeconomic and demographic attributes are grouped together; and also 2) association: each non-sampled county holds a value of the mortality rate of the sampled county that is most similar to itself.

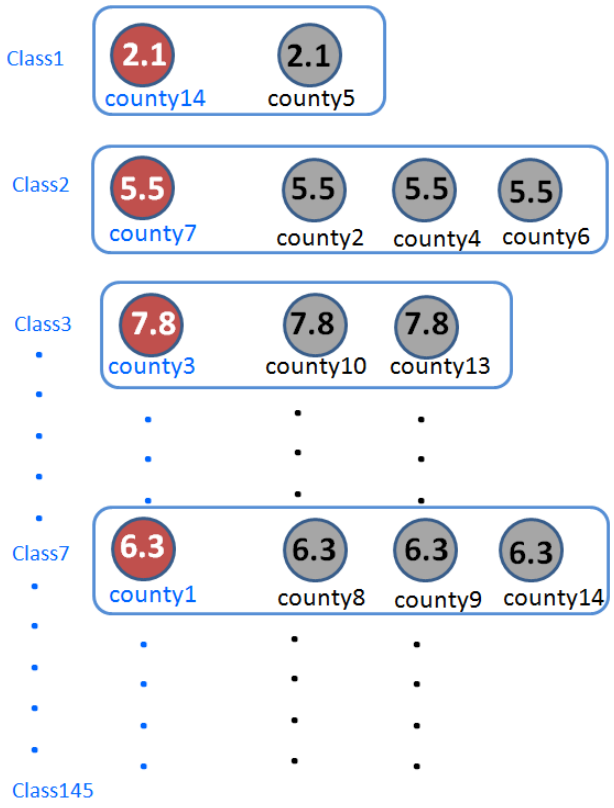
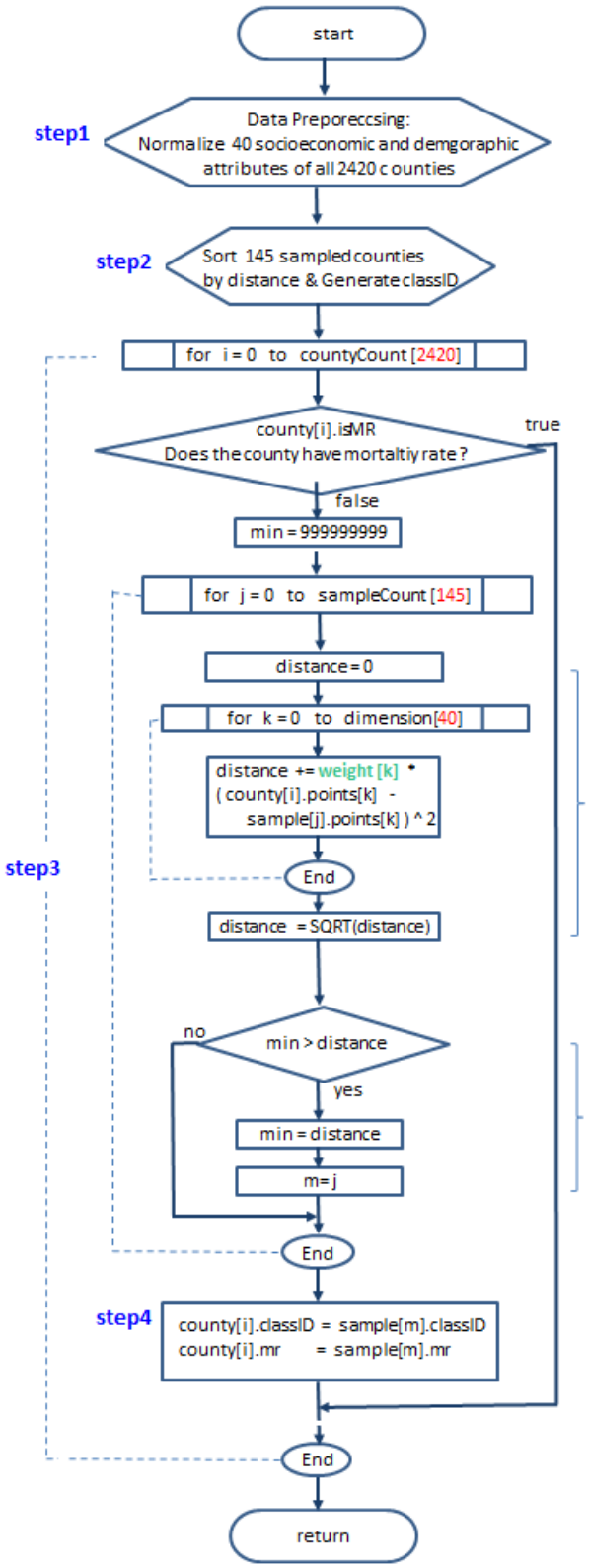


Figure 34

Output of SSC: Red nodes represent sampled counties and grey nodes represent non-sampled counties. A value of each node represents a mortality rate. Socioeconomically and demographically similar counties are grouped together in each class, and counties in neighboring classes are more similar than counties belonging to non-neighboring classes. Each class always contains only one sampled county.

The detailed algorithm for steps 3 and 4 is illustrated in flowchart and pseudo C language (Figure 35). The entire program is available in Appendix B.




```
// All Counties
typedef struct{
   int   countyID;      // input
   int   dataID;        // input
   int   isMR;          // output
   int   classID;       // output
   float distance;      // output
   float mr;            //output
   float *points;       // 40 socioeconomic/demogrpahic input
} county[2420];
countyCount= 2420;
```

```
// Sampled Counties
typedef struct{
   int   countyID;      // input
   int   classID;       // output
   float mr;            // input
   float *points;       // 40 socioeconomic/demographic input
} sample[145];
sampleCount = 145;
```

Get distance between county[i].points and sample[j].points

In this section, we assume that all variables have equal weights (impacts) on the similarity measure. (weight = 1) . Please see section 4.3.6 for setting different weights.

The variable m holds the index of best matching unit (BMU)

Figure 35 (cont. on next page)

```
step1   00 Data preprocessing: normalize socioeconomic and demogrpahic attributes
step2   01 Sort sampled counties by distance. After the sorting, class ID is generated
        02 for i=0 to countyCounty//the number of non-sampled counties
        03   min = 999999999
        04   for j=0 to sampleCount//the number of sampled counties
        05     distance = 0
step3   06     for k=0 to attributeCount//the number of socioeconomic and demographic attributes
        07       distance += weight[k] * ( county[i].points[k] – sample[j].points[k] )^2
        08     end for k
        09     distance = sqrt(distance)
        10     if (min > distance ) {
        11       min = distance
        12       m=j  // index of BMU
        13     }
        14   end for j
        15   county[i].classID   = sample[m].classID;
        16   county[i].mr        = sample[m].mr;          step4
        17 end for i
```

Figure 35 The SSC algorithm.

Step by step description of the SSC method in the text corresponds to the steps described in the flow chart and pseudo code.

The first output of the SSC algorithm is 145 classes containing all of the counties in the mainland of China. Each of the 145 classes holds counties having similar population characteristics, and those counties that belong to classes next to one another are similar to one another. For example, counties in class2 are more socioeconomically and demographically similar to counties in class1 or counties in class3 rather than counties in class7 in Figure 34. Regarding the visualization of SSC results, class *id* of each county is mapped (Figure 36).

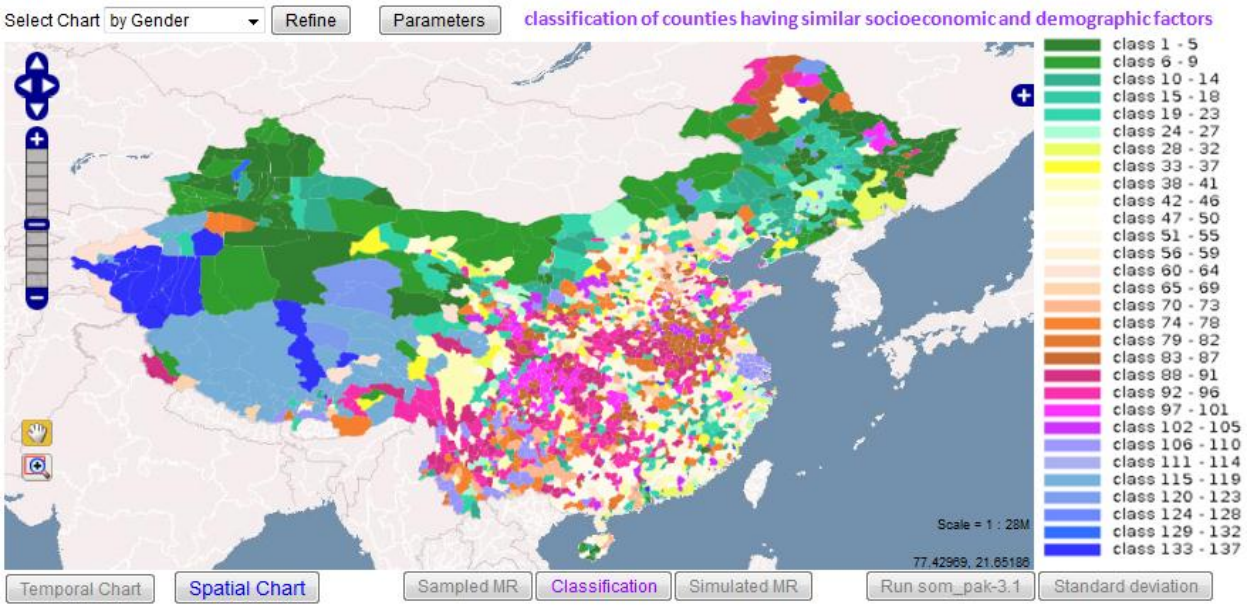


Figure 36

An output of SSC: class *id* of each county is visualized to represent the patterns of socioeconomically and demographically similar counties. The other output of SSC is shown in Figure 39.

Since counties belonging to classes next to one another mean that they are similar to one another, counties belonging to neighboring classes are represented by the same or similar colors. For the visualization purpose, 145 classes are grouped into 30 colors from the first class to the last class are designed to change gradually to represent the similarity of counties belonging to nearby classes and the dissimilarity of counties belonging to classes separated far away. For example, since classes 1 through 5 are similar to one another, they are represented by the same color, the dark green. A group of classes from 6 to 9 is similar to a group of classes from 1 to 5 and from 10 to 14. Thus, all the counties belonging to all these classes from 1 through 14 are represented in similar colors, dark to medium green colors on the map. One the other hand, all the greenish colored counties are socioeconomically and demographically different from reddish colored

counties that belong to the classes far away from the greenish colored classes. Figure 37 shows the clusters of socioeconomically and demographically similar counties.

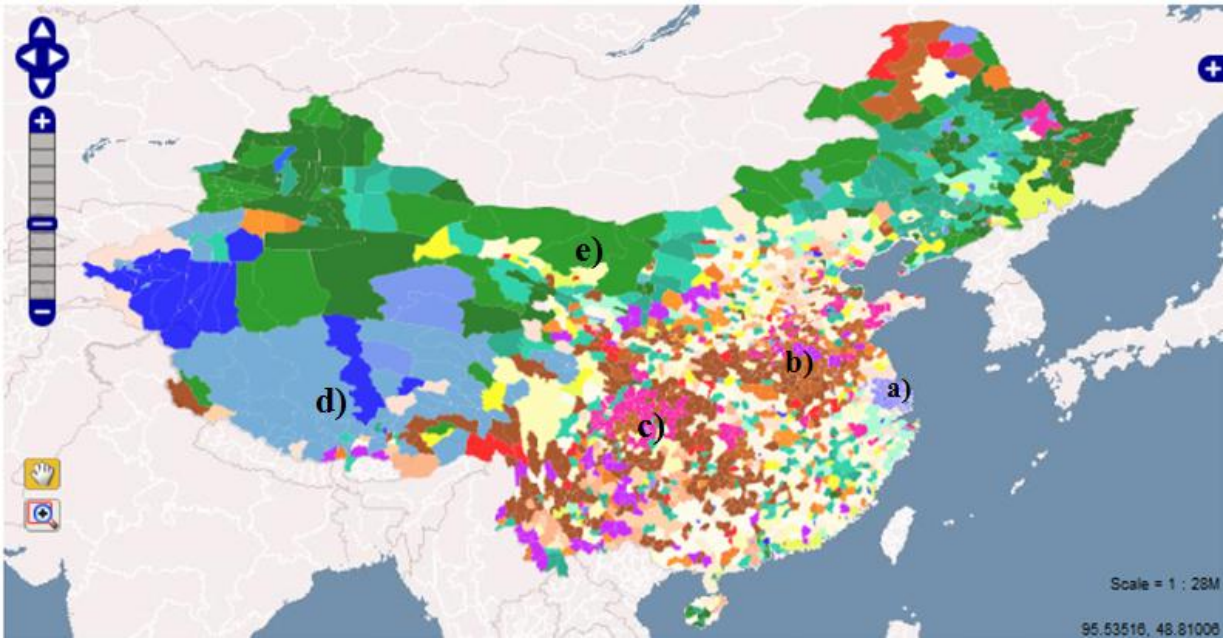


Figure 37

Major clusters of socioeconomically and demographically similar counties.

a) cluster: light purple colored counties on the eastern coastal areas
b) cluster: brown colored counties in inner-land areas of east coast
c) cluster: pink colored counties in the center region
d) cluster: blue colored counties in southwest areas
e) cluster: green colored counties across northwest areas to northeast areas

The second output of the SSC algorithm is the estimated mortality rates of non-sampled counties. The visualization of the mortality rate of each county in the geographic space from the SSC output (described in Figure 34) reveals the patterns of the estimated mortality rates of the entire China (mainland). Figure 39 shows a set of example maps of the distributional patterns of estimated mortality rates that are derived by using the SSC as part of the CyberGIS application. Many types of such maps as for mortality rates specific to gender, age and death causes, can be generated by user's selections of input parameters through the CyberGIS application interface

(see section 3.2.1 for details). Maps in Figure 38 are the visualization of the mortality rates of sampled counties (i.e. inputs of SSC method). With the two input data of the mortality rates of sampled counties (Figure38) and socioeconomic and demographic data (Table1), SSC method reveals the distributional patterns of estimated mortality rates that are shown in Figure39. A color ramp in which colors change gradually from blue to red is created to represent the areas with relatively low or high mortality rates.

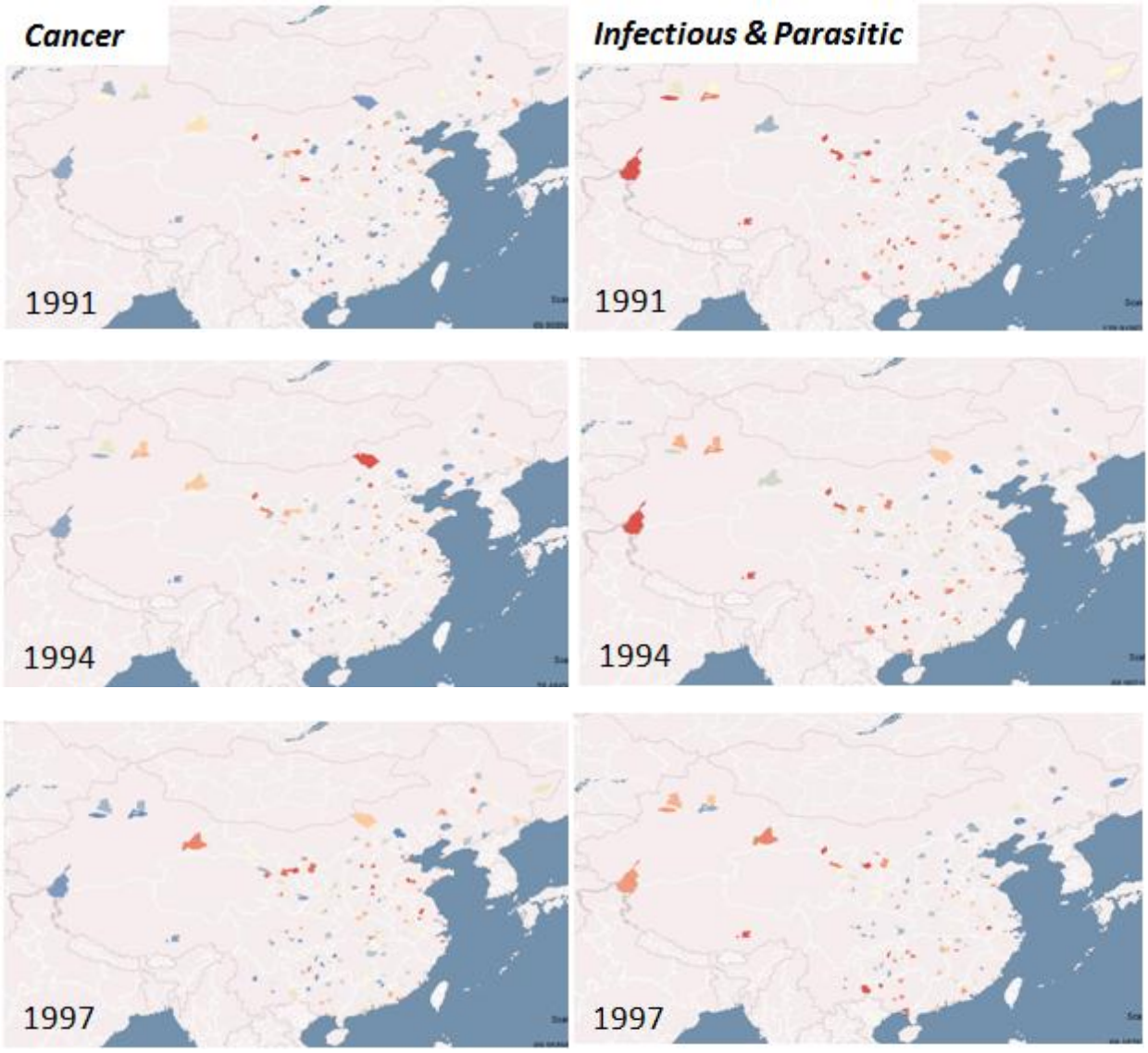


Figure 38 (cont. on next page)

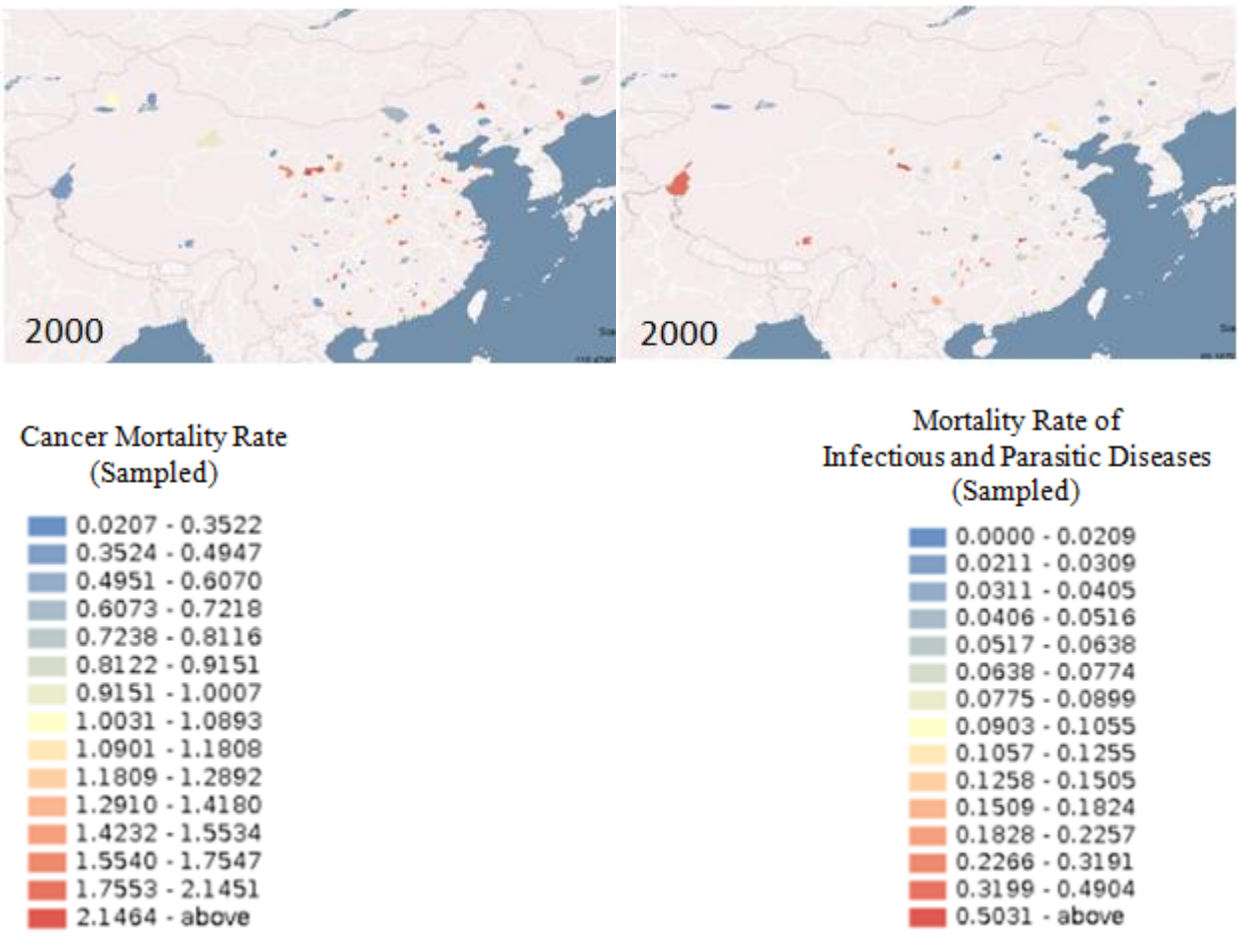


Figure 38

The visualization of the mortality rates of sampled counties. These are originally available data and also the input of SSC method. Maps on the left represent the mortality rates of cancer of sampled counties. Maps on the right show the mortality rates of infectious and parasitic diseases of sampled counties.

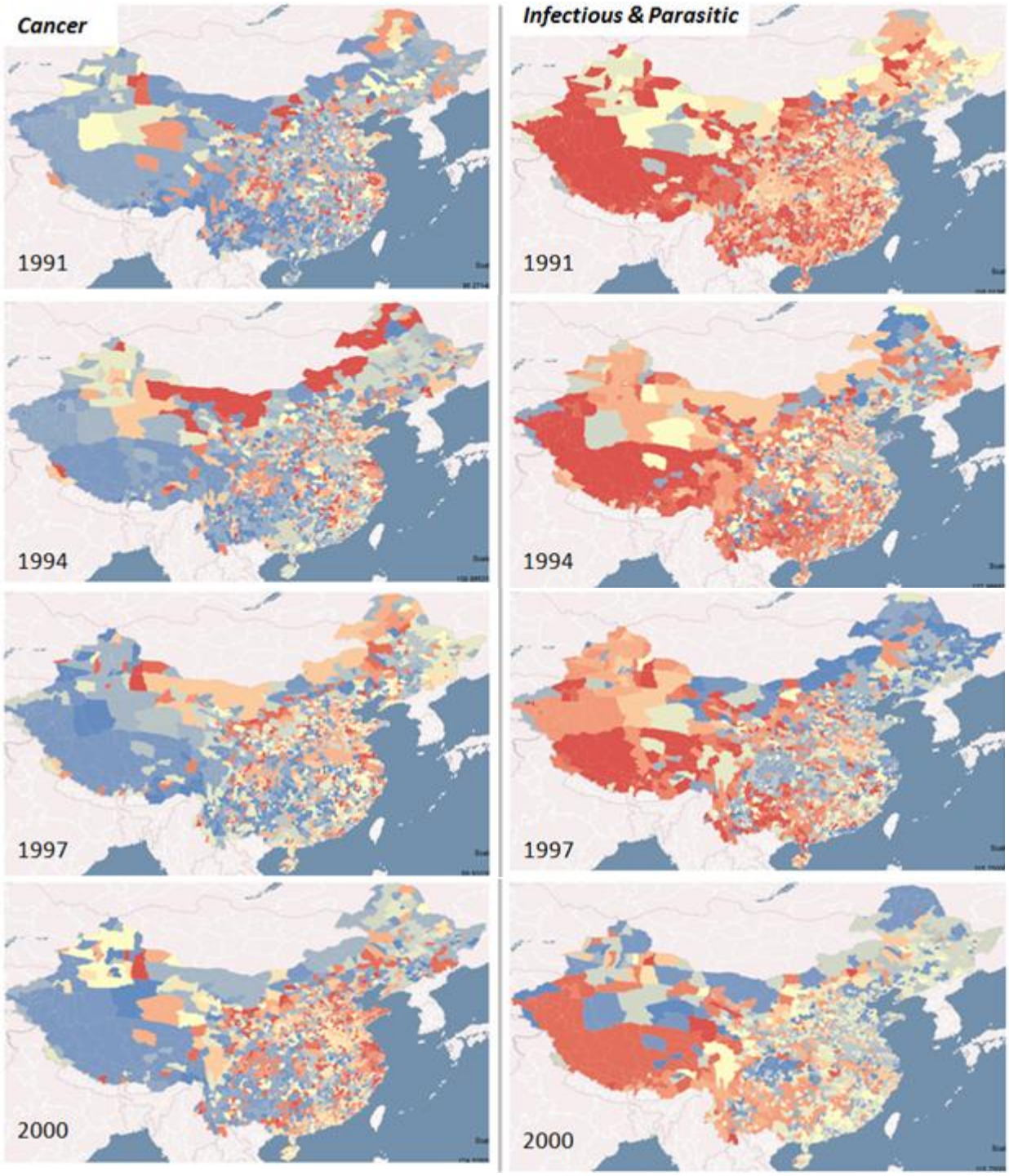


Figure 39 (cont. on next page)

Figure 39

The visualization of the estimated mortality rates of non-sampled counties and known mortality rates of sampled counties. The estimated mortality rates of non-sampled counties are one of the outputs of the SSC method. Maps on the left represent the estimated mortality rates of cancer. Maps on the right show the estimated mortality rates of infectious and parasitic diseases.

To interpret the estimated mortality rates on the maps, it makes sense to focus on the overall understanding of the distributional patterns of high or low mortality rates rather than picking up the exact value of the mortality rates of specific non-sampled counties. The actual mortality rate of a non-sampled county is originally unknown. However, through the analysis using the SSC method, one of the sampled counties is selected and a mortality rate of the sampled county is assigned to the non-sampled counties that are most socioeconomically and demographically similar to the chosen sampled county. As a result, a value of mortality rate of each non-sampled county on the map is equal to a mortality rate of one of the sampled counties. In this way, the mortality rates of non-sampled counties are estimated rather than being predicted. Therefore, though each of the 15 classes might indicate the exact range of mortality rates based on the legend of Figure 39, it may not be appropriate to interpret those values of the mortality rates on the maps are exactly within the range of a particular class among the 15 classes.

Though the SSC method developed based on the exploratory data analysis cannot predict the exact values of mortality rates of specific non-sampled counties, it provides a hint about the areas with relatively high or low mortality rates. For example, the 4th class of cancer mortality rate in the legend of Figure 39 shows that the mortality rate is between 0.6 and 0.72. Under the assumption that the mortality rate tends to be similar within the socioeconomically and demographically similar counties, the actual mortality rate represented by the 4th class will likely be around 0.6 or 0.7. The assumption makes sense when it is considered that the 145 sampled counties are "nationally representative samples reflected regional population distributions, urban and rural areas, age and sex, and eastern, middle, and western regions" (Yang *et al*. 2008). It implies that with only 145 sampled counties, it is possible to capture the mortality trends of overall population since the non-sampled counties are similar to some of the 145 sampled counties. Furthermore, each mortality rate of a non-sampled county can be represented by one of the mortality rates of a sampled county when the population characteristics of the sampled county are similar to the population characteristics of the non-sampled county.

The mortality rates of cancer and infectious/parasitic diseases show different patterns (Figure 39). Since chronic diseases had dramatically increased during the 10-year period (Yang *et al*. 2008), cancer as one of the prevalent death causes among chronic diseases is chosen to be mapped for this case study. In addition, the patterns of the mortality rates of chronic diseases are compared and contrasted with the patterns of the mortality rates of infectious and parasitic diseases. Cancer mortality rates increased especially along the east coastal areas of China. Especially, two maps of the cancer mortality rates of 1991 and 2000 show a dramatic increase in terms of the areas of high cancer mortality rates. On the other hand, the southwest part of China consistently shows low cancer mortality rates. In contrast, the mortality rates of infectious and

parasitic diseases decreased during the 10 years, especially along the east coastal areas and the central region of China. Comparing the two maps of 1991 and 2000, the mortality rates of infectious and parasitic diseases show a significant decrease across the entire country. However, the mortality rates of the west part of China remained relatively high compared to other regions for the 10 years. Especially, the comparison of the two maps of mortality rates of cancer and infectious diseases of 2000 shows the reverse distributional pattern (Figure 40)—i.e. the mortality rates of cancer tend to be high in the areas where the mortality rates of infectious and parasitic diseases are low whereas the mortality rates of cancer seem to be low in the areas where the mortality rates of infectious and parasitic diseases are high.

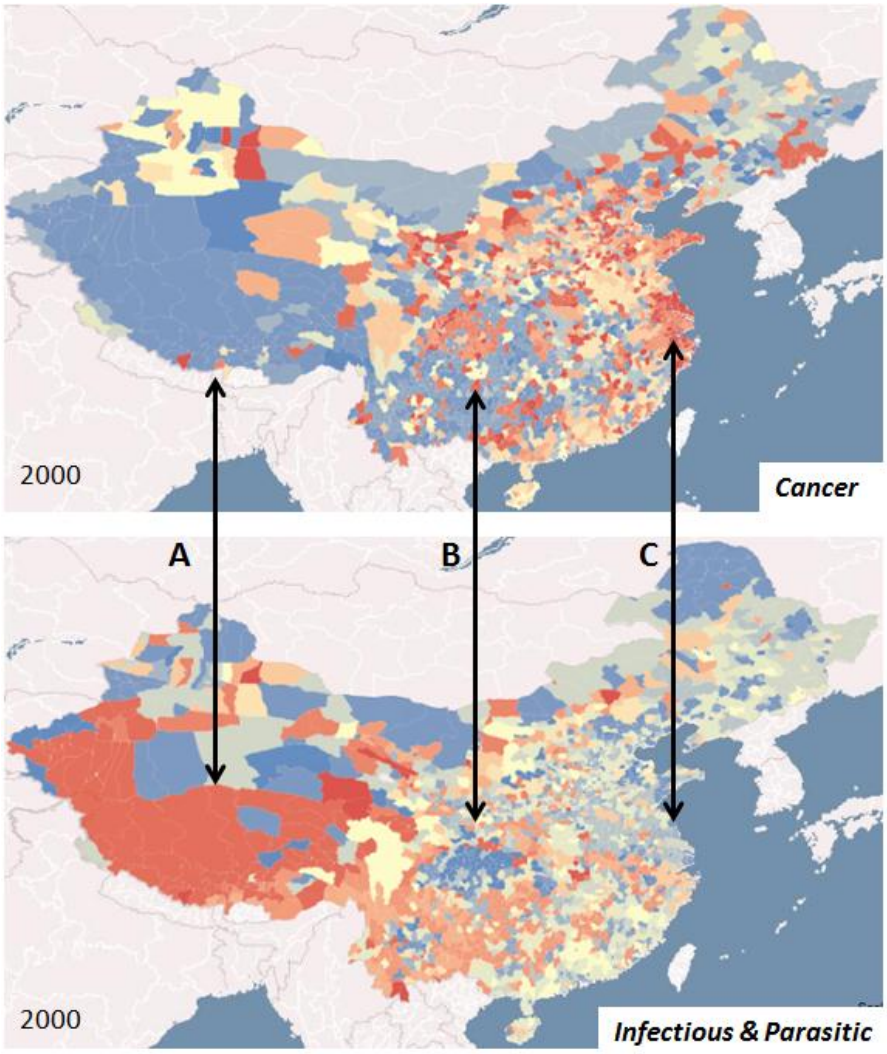


Figure 40

Contrasting the distributional patterns of the mortality rates of cancer and the mortality rates of infectious and parasitic diseases: especially the southwest (A), central south (B), and eastern coastal (C) regions show a contrast in the mortality rates of the two types of diseases.

4.3.6 Weighting and spatial factors of straightforward spatial classification (SSC)

A user can assign a weight to each of the 40 socioeconomic and demographic variables (Table1) so that each variable can be evaluated in terms of its specified level of impact on the similarity measure. In the section 4.3.4 and 4.3.5 we assume that all variables have equal weight on the similarity measure (see flowchart in Figure 35). On the interface of the CyberGIS application the user has an option to type in different weight for each variable and, thus, can create maps similar to maps in figure 36 and 39 with different weight of each socioeconomic and demographic variable. When the user clicks "Parameters" button on the top of the interface of Figure11, the option of parameter setting shows up (Figure 41). Weight for each variable is set to 1 by default, but the user can change it as needed. For example, the user may think that the mortality rates are likely to be similar in the areas where the percentage of aging people is similar. Therefore, the user would think that particular variables (e.g. age is 65 and older) should have more impact on measuring the similarity of counties than any other variables. In this case, the user can assign more weight (i.e. greater than 1) to the old population.

X, Y coordinates of centroid at the bottom right of Figure 41 has x, y coordinate of centroid of each county. Increasing weight to the x, y coordinate of centroid means country location has more impact on the similarity measure than other 40 socioeconomic and demographic variables. Therefore, straightforward spatial classification (SSC) method with this increased weighting on x, y coordinates of centroid produces the results that the socioeconomic and demographic attributes and mortality rates are more similar in counties geographically neighboring to one another than in faraway counties. The larger the weight of x, y coordinates is, the more similar the socioeconomic and demographic attributes and the mortality rates of neighboring counties are.

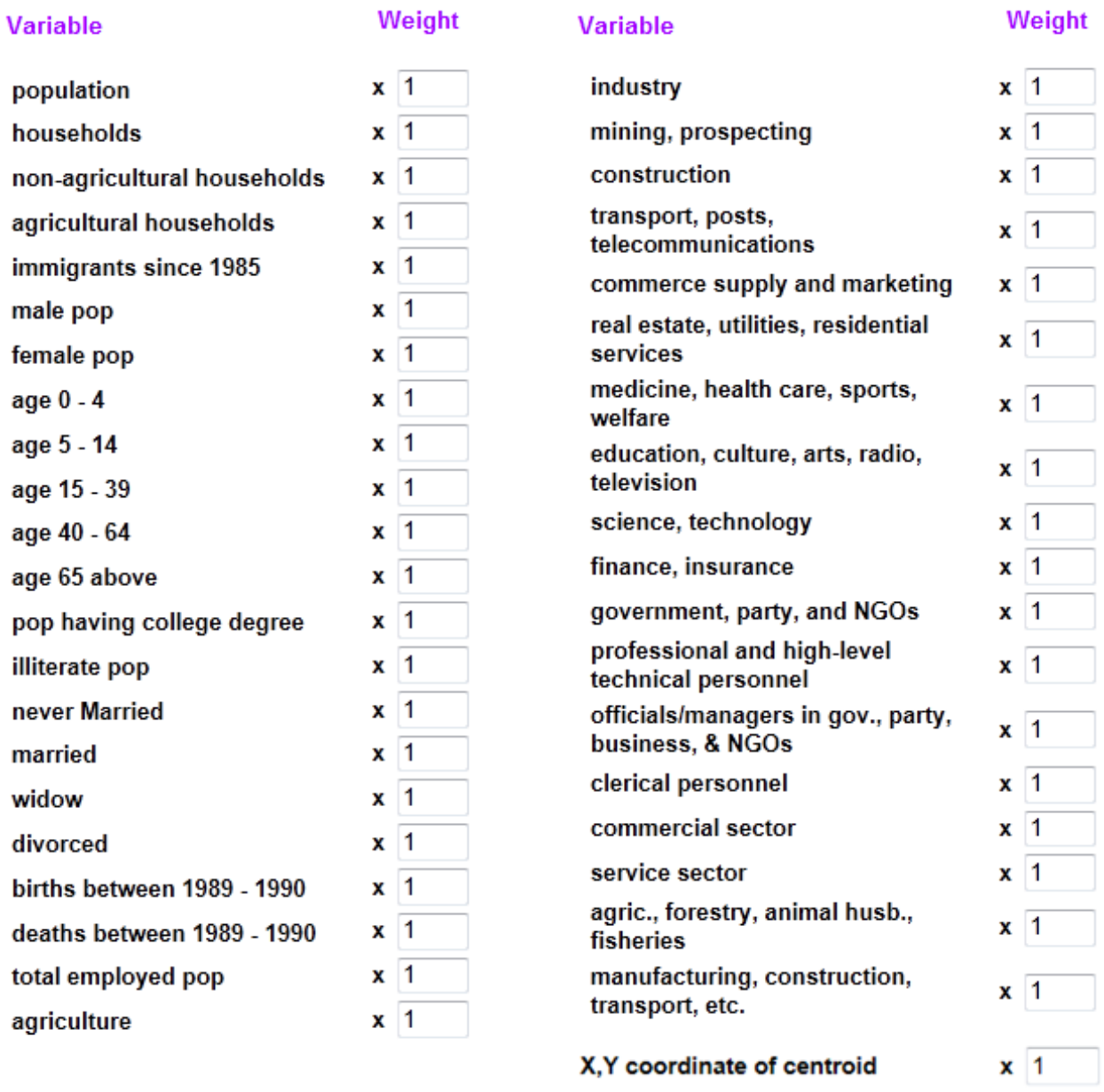


Figure 41
User's option to set different weight to each variable

Figure 42 represents that different weights of x,y coordinates produce different results of the SSC method - e.g. the classification of socioeconomic and demographic attributes of all counties and the estimated mortality rates of 1991, 1994, 1997 and 2000. Five maps in the first column are the output of SSC when the weight of x, y coordinate is equal to 0. Weighting 0 on x,

y coordinates means that spatial factors are not explicitly considered in producing those maps – e.g. the mortality rate of a county is not directly affected by the mortality rates of neighboring counties. In the second case (five maps in the middle column), weight 1 is assigned to x, y coordinates, which means that the variable of x, y coordinates has the same-level impact on measuring the similarity of counties as much as the other 40 socioeconomic and demographic variables. Also, by weighting 1 to the x, y coordinates of centroid, the mortality rates of counties are affected by the mortality rates of neighboring counties. In other words, the location of counties is one of the factors that we use to estimate the mortality rates of counties. When maps in the first column (weight 0 on x, y coordinates) and the second column (weight 1 on x, y coordinates) are compared, there is a general trend that the socioeconomic and demographic attributes and the mortality rates of neighboring counties are more similar in the maps of the second column than in the maps in the first column.

In the third case (five maps in the last column), a considerably high weight value - 40 – is assigned to x, y coordinates of centroid to experiment the effect of such weighting. In this case, the SSC method produces an output where the effect of the locations of counties is a dominant factor in measuring the similarity of socioeconomic and demographic attributes and measuring the mortality rates. The socioeconomic and demographic attributes are similar in neighboring counties and also the counties having the same values of mortality rates are spatially aggregated in neighboring counties. Even though the third case is not realistic in estimating the mortality rates, the comparison of output maps of the first, the second and the third cases shows the influence of increasing weight on x, y coordinates of centroid. This further demonstrates the flexibility and capability of SSC for taking into account geographic distance effects.

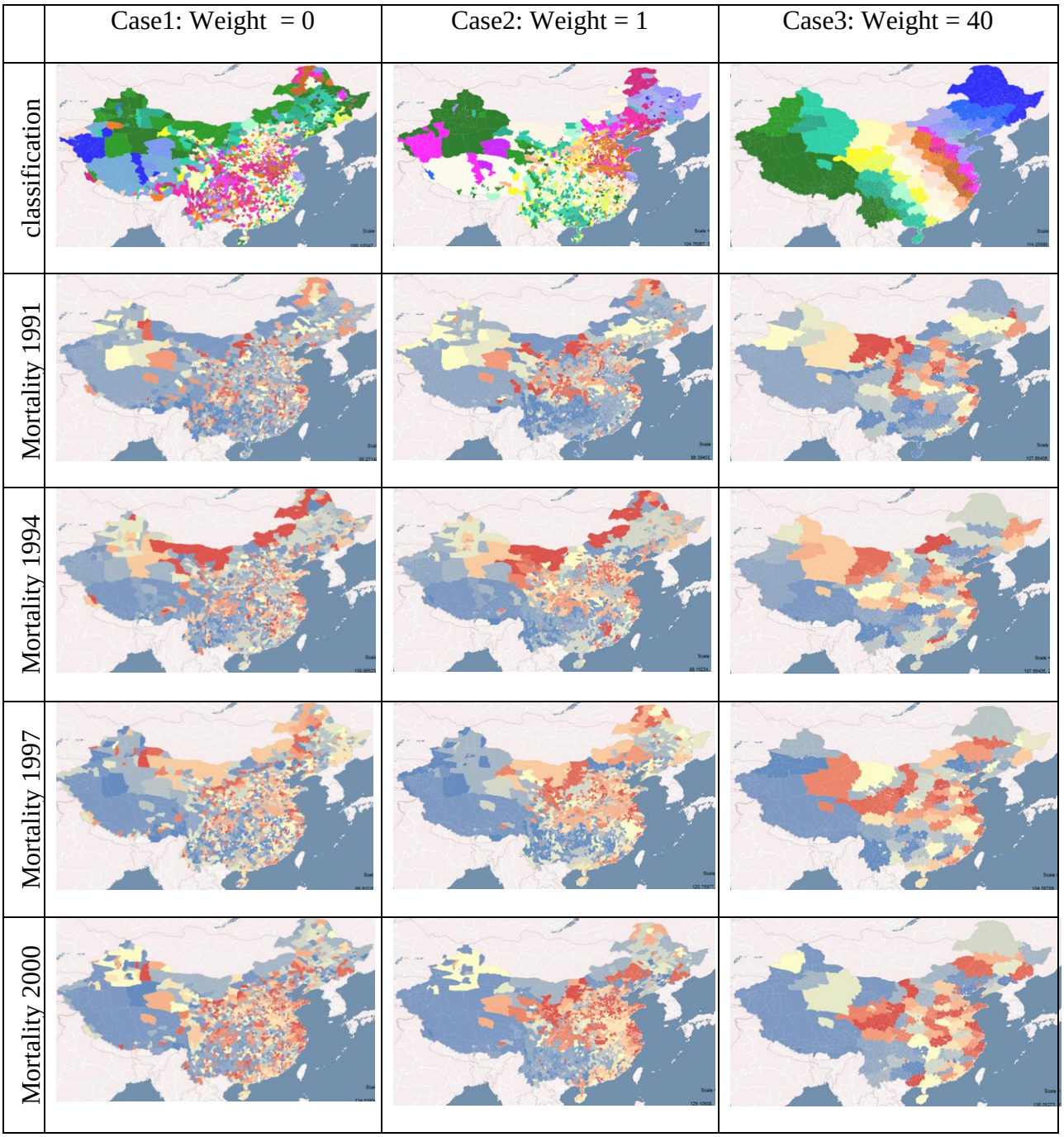


Figure 42 (cont. on next page)

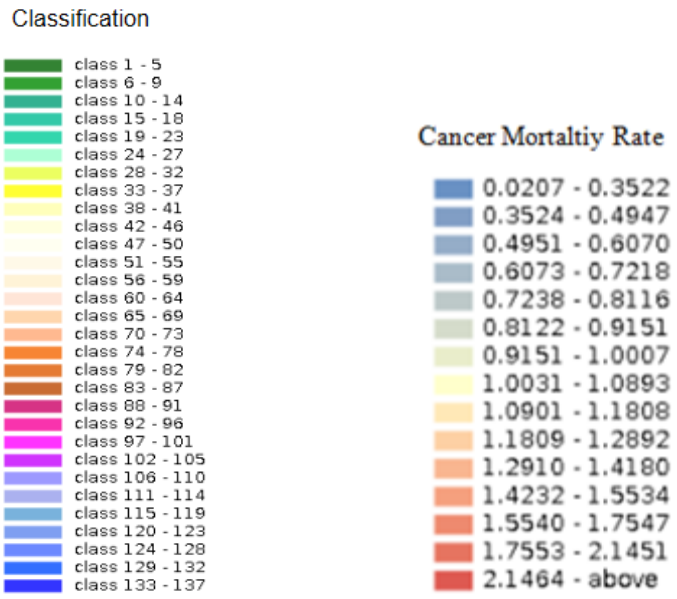


Figure 42
These maps are the results of the SSC method with different options of weighting x,y coordinates of centroid. SSC with weight 0,1 and 40 of x, y coordinates produces each 5 maps on the left, in the middle, and on the right. As the weight of x,y coordinates, both the socioeconomic and demographic attributes and the estimated mortality rates are likely to show more similarities in geographically neighboring counties than in faraway ones.

CHAPTER 5

SUMMARY AND CONCLUDING DISCUSSION

5.1 SUMMARY OF FINDINGS

The integrated approach to combining the self-organizing map (SOM) and straightforward spatial classification (SSC) methods effectively addresses the challenge of analyzing the sparsely sampled mortality data and high-dimensional socioeconomic and demographic data. The exploratory data analysis based on this approach leads to new knowledge and insights for (1) identifying areas that should be sampled to have better coverage reflecting all of the population characteristics, (2) assessing the effectiveness of the national sampling strategy, and (3) evaluating the relationships between mortality rates and socioeconomic and demographic factors. Both SOM and SSC use Euclidean distance to measure the similarity of high-dimensional data in attribute space—i.e. socioeconomic and demographic data consisting of 40 attributes. The SSC has demonstrated that the mortality rates of the non-sampled counties can be represented by the mortality rate of the sampled counties based on the hypothetical similarity of socioeconomic and demographic factors among counties. The SSC serves to understand the spatial patterns of the national mortality by using the sparsely sampled mortality data and to identify the regions that have the relatively high mortality rates compared to those in other regions.

The case study demonstrates the findings based on the maps generated by the spatial analyses supported in the CyberGIS application. One of the findings that can be examined by using the application is that the cancer mortality rate had increased especially along the east coastal areas of China from 1991 to 2000. When the pattern maps of the mortality rates of cancers are compared and contrasted with those of parasitic and infectious diseases, the mortality

rates of the two types of diseases have shown the spatially reversed patterns. Beyond this finding, CyberGIS application enables users to temporally explore the patterns of the mortality rates of any death causes among the population of different gender and age groups.

The interface design and service components of the CyberGIS application have demonstrated how the integrated use of visualization tools with the spatial analyses can contribute to the exploratory analysis of data in various settings. A possible scenario is that users first examine an overview of the patterns of the mortality rates of the entire China while visual analytical tools are used to reveal distributional characteristics of the mortality rates at the provincial level. Once some areas of interest are identified from the overview, users can easily drill down to examine the distributional characteristics at the sampled county within the areas of interest.

5.2 Limitation

There are limitations in analyses that are conducted by SOM and SSC methods based on exploratory data analysis. This study focuses on exploring available datasets, which help to generate hypotheses rather than hypothesis testing. Generating new hypotheses based on the exploration of the data that can be easily done with the CyberGIS application would be able to lead to the next level of analysis such as modeling to predict the mortality rates of non-sampled locations. However, this study does not produce a confirmatory result in predicting them. Instead, this study is focused on exploring various maps with various options of input parameters. In other words, SOM and SSC methods produce various maps in measuring the similarity of counties and estimating the mortality rates rather than any confirmatory result. The user is required to set the various variables to run the SOM and SSC methods and generate result maps of those spatial analyses. For example, there are many options that the user needs to specify such

as the number of the nodes within SOM and weight of each SSC variable. Therefore, the maps produced as the outputs of the spatial analyses may change depending on the user's settings of the variables. In other words, it is the user's choice to decide which variables he/she would set. The resulting maps in the case study of this thesis are chosen after many experiments with different variable options, and ones that can be considered producing the most meaningful and helpful results to solve the problem are picked.

5.3 BENEFITS OF CYBERGIS IN PUBLIC HEALTH

Spatial analysis of public health data supported by conventional GIS often involves complicated steps, and may require in-depth technical training. Specifically, the following issues may arise. Users' data have many variables and they need to selectively query and configure large amounts of input data with different combinations of variables. Thus, users may have to write or understand query commands for data selection, produce many input files, and convert them to appropriate formats for the preparation of spatial analysis. In addition, manual processes of data preprocessing such as normalization or standardization are often involved. Sometimes, the same process of data preprocessing, formatting, importing of input files, setting up input parameters, map unit, and extent, coloring maps, creating legends, and exporting result maps has to be repeated by spatial analysts for many times and, thus, may become error-prone. When there are needs to compare and contrast result maps of a spatial analysis for a certain period of time, e.g. 10 years with every month considered, the whole process may become very inefficient and labor intensive.

The integration of the spatial analyses, mortality data, and highly interactive user interface with a set of visualization tools into the CyberGIS environment has this process optimized. Specifically, users query input data with a few mouse clicks through the user-friendly

interface without the involvement of typing query commands, and selected input data are automatically available for data preprocessing, spatial analyses, and visualization. Furthermore, output maps can be seen and are compared easily with a few button clicks on the fly without exporting and generating an image file of result maps. In addition, interactive visualization tools are used simultaneously with dynamically created maps based on the analyses. Therefore, the CyberGIS application has dramatically reduced the time required for data processing and performing spatial analysis. Furthermore, it allows public health users including non-GIS experts to perform spatial analysis with minimal effort and effectively engage spatial analytics. Therefore, the CyberGIS environment has a potential to increase the efficiency of spatial data analytics in public health research and practice.

In addition, the limitations of conventional GIS in its computational scalability have been addressed by integrating cloud computing based on user-centered service-oriented architecture. The user-centered interface guides users to access powerful cyberinfrastructure resources, which allows a number of users to run their spatial analyses simultaneously while protecting sensitive public health data. As a future work, the integration of the spatial analysis services and cloud infrastructure can be further extended to include other CyberGIS resources and services.

REFERENCES

Ahmad, O. B., C. Boschi-Pinto, A. D. Lopez, C. J. L. Murray, R. Lozano, and M. Inoue. 2001. Age Standardization of Rates: a new WHO standard. GPE Discussion Paper Series: No. 31. Geneva, World Health Organization. http://www.who.int/healthinfo/paper31.pdf (last accessed 6 Dec 2012).

Agarwal, P., and A. Skupin. 2008. *Self-organising maps: Applications in geographic information science.* John Wiley & Sons Inc.

Anderson, G., and R. Moreno-Sanchez. 2003. Building web-based spatial information solutions around open specifications and open source software. *Transactions in GIS* 7 (4): 447-66.

Anselin, L., I. Syabri, and Y. Kho. 2006. GeoDa: An introduction to spatial data analysis. *Geographical Analysis* 38 (1): 5-22.

Armstrong, M. P., G. Rushton, and D. L. Zimmerman. 1999. Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18 (5): 497-525.

Atkins, D. 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure.

Basara, H. G., and M. Yuan. 2008. Community health assessment using self-organizing maps and geographic information systems. *International Journal of Health Geographics* 7: 67.

Blanton, J. D., A. Manangan, J. Manangan, C. A. Hanlon, D. Slate, and C. E. Rupprecht. 2006. Development of a GIS-based, real-time internet mapping tool for rabies surveillance. *International Journal of Health Geographics* 5 (1): 47.

Boulos, M. N. K., and K. Honda. 2006. Web GIS in practice IV: Publishing your health maps and connecting to remote WMS sources using the open source UMN MapServer and DM solutions MapLab. *International Journal of Health Geographics* 5 (1): 6.

Buja, A., D. Cook, and D. F. Swayne. 1996. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*: 78-99.

Cleveland, W.S., 1993, *Visualizing Data* (Summit, NJ: Hobart Press).

Cromley, E. K. 2003. GIS and disease. *Annual Review of Public Health* 24 (1): 7-24.

Cromley, E. K., and S. L. McLafferty. 2012. *GIS and Public Health, 2nd Edition.* The Guilford Press.

Croner, C. M. 2004. Public health GIS and the internet. *Journal of Map and Geography Libraries* 1 (1): 105-35.

———. 2003. Public health, GIS, and the Internet 1. *Annual Review of Public Health* 24 (1): 57-82.

Curtin, L. R., R. J. Klein, and National Center for Health Statistics (US). 1995. *Direct standardization (age-adjusted death rates)* US Dept. of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

National Science Foundation (NSF). 2007. *Cyberinfrastructure vision for 21st century discovery*. National Science Foundation, Cyberinfrastructure Council. http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf (last accessed 6 Dec 2012).

Dailey, G. 2006. Normalizing census data using ArcMap. *ArcUser Online, January-March 2006*: 52-3.

Dangermond, J. 2002. Web services and GIS. *Geospatial Solutions* 12 (7): 56-7.

Evans, B., and C. E. Sabel. 2012. Open-source web-based geographical information system for health exposure assessment. *International Journal of Health Geographics* 11 (1): 2.

Good, I. J. 1983. The philosophy of exploratory data analysis. *Philosophy of Science*: 283-95.

Gould, P. 1981. Letting the data speak for themselves. *Annals of the Association of American Geographers* 71 (2): 166-76.

Guo, D. 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science* 22 (7): 801-23.

Guo, D., M. Gahegan, A. M. MacEachren, and B. Zhou. 2005. Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science* 32 (2): 113.

Guo, D., and H. Jin. 2011. iRedistrict: Geovisual analytics for redistricting optimization. *Journal of Visual Languages & Computing*.

Hanafi-Bojd, AA, H. Vatandoost, MA Oshaghi, Z. Charrahy, AA Haghdoost, G. Zamani, F. Abedi, MM Sedaghat, M. Soltani, and M. Shahi. 2012. Spatial analysis and mapping of malaria risk in an endemic area, south of iran: A GIS based decision making for planning of control. *Acta Tropica*.

Harris, R. L. 1999. *Information graphics: A comprehensive illustrated reference.* Oxford University Press, USA.

Hecht, L. 2002. Open GIS connection insist on interoperability! *Geo World* 15: 22-3.

Kamadjeu, R., and H. Tolentino. 2006. Web-based public health geographic information systems for resources-constrained environment using scalable vector graphics technology: A proof of concept applied to the expanded program on immunization data. *International Journal of Health Geographics* 5 (1): 24.

Kaski, S., and T. Kohonen. 1996. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. Paper presented at Neural Networks in Financial Engineering. *Proceedings of the Third International Conference on Neural Networks in the Capital Markets* 498-507.

Ki, S. J., J. H. Kang, S. W. Lee, Y. S. Lee, K. H. Cho, K. G. An, and J. H. Kim. 2011. Advancing assessment and design of stormwater monitoring programs using a self-organizing map: Characterization of trace metal concentration profiles in stormwater runoff. *Water Research 45:4183-4197*.

Kohonen, T., J. Hynninen, J. Kangas, and J. Laaksonen. 1996. Som pak: The self-organizing map program package. *Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science*.

Kohonen, T., MR Schroeder, TS Huang. 2001. *Self-Organizing Maps.* Springer-verlag New York. *Inc., Secaucus, NJ*.

Koua, E. L., and M. J. Kraak. 2004. A usability framework for the design and evaluation of an exploratory geovisualization environment. *Proceedings of the 8th international conference on information visualization, IEEE Computer Society Press* 153-158

MacEachren, A., S. Crawford, M. Akella, and G. Lengerich. 2008. Design and implementation of a model, web-based, GIS-enabled cancer atlas. *Cartographic Journal, the* 45 (4): 246-60.

Maclachlan, J. C., M. Jerrett, T. Abernathy, M. Sears, and M. J. Bunch. 2007. Mapping health on the Internet: A new tool for environmental justice and public health research. *Health & Place* 13 (1): 72-86.

Mehmood, Y., M. Abbas, X. Chen, and T. Honkela. 2011. Self-organizing maps of nutrition, lifestyle and health situation in the world. *Advances in Self-Organizing Maps*: 160-7.

Messner, S. F., L. Anselin, R. D. Baller, D. F. Hawkins, G. Deane, and S. E. Tolnay. 1999. The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. *Journal of Quantitative Criminology* 15 (4): 423-50.

Moreno-Sanchez, R., G. Anderson, J. Cruz, and M. Hayden. 2007. The potential for the use of open source software and open specifications in creating Web-based cross-border health spatial information systems. *International Journal of Geographical Information Science* 21 (10): 1135-63.

Oyana, T. J., J. Yan, and J. S. Lwebuga-mukasa. 2005. Exploration of geographic information systems-based medical databases with self-organizing maps: A case study of adult asthma. In Proceedings of the 8th International Conference on GeoComputation, August 1st-3rd, 2005, Ann Arbor, University of Michigan http://www.geocomputation.org/2005/Oyana.pdf (last accessed 6 Dec 2012).

Páez, A., J. Gallo, R. N. Buliung, and S. Dall'Erba. 2010. Progress in spatial analysis: Introduction. *Progress in Spatial Analysis*: 1-13.

Pirotti, F., A. Guarnieri, and A. Vettore. 2011. Collaborative Web-GIS design: A case study for road risk analysis and monitoring. *Transactions in GIS* 15 (2): 213-26.

Richards, T. B., C. M. Croner, G. Rushton, C. K. Brown, and L. Fowler. 1999. Geographic information systems and public health: Mapping the future. *Public health reports-US* 114 : 359-73.

Rinner, C., B. Moldofsky, M. D. Cusimano, S. Marshall, and T. Hernandez. 2011. Exploring the boundaries of web map services: The example of the online injury atlas for Ontario. *Transactions in GIS* 15 (2): 129-45.

Robertson, C., and T. A. Nelson. 2010. Review of software for space-time disease surveillance. *International Journal of Health Geographics* 9 (1): 16.

Rop, M., Y. Liu, and M. C. Wimberly. FWA–A framework for developing web-atlas applications.

Skupin, A., and P. Agarwal. 2008. Introduction: What is a Self-Organizing map? *Self-Organizing Maps*: 1-20.

Supak, S., H. Luo, L. Tateosian, K. Fang, J. Harrell, C. Harrelson, A. D. Bailey, and H. Devine. 2012. Who's watching your food? A flexible framework for public health Monitoring1. *Transactions in GIS* 16 (2): 89-104.

Tiwari, C., and G. Rushton. 2010. A spatial analysis system for integrating data, methods and models on environmental risks and health outcomes. *Transactions in GIS* 14: 177-95.

Törönen, P., M. Kolehmainen, G. Wong, and E. Castrén. 1999. Analysis of gene expression data using self-organizing maps. *FEBS Letters* 451 (2): 142-6.

Toutant, S., P. Gosselin, D. Bélanger, R. Bustinza, and S. Rivest. 2011. An open source web application for the surveillance and prevention of the impacts on public health of extreme meteorological events: The SUPREME system. *International Journal of Health Geographics* 10 (1): 39.

Valkonen, V. P., M. Kolehmainen, H. M. Lakka, and J. T. Salonen. 2002. Insulin resistance syndrome revisited: Application of self-organizing maps. *International Journal of Epidemiology* 31 (4): 864-71.

Vanmeulebrouk, B., U. Rivett, A. Ricketts, and M. Loudon. 2008. International journal of health geographics. *International Journal of Health Geographics* 7 : 53.

Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas. 1999. Self-organizing map in matlab: The SOM toolbox. Paper presented at Proceedings of the Matlab DSP Conference, .

Wang, S. 2010. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers* 100 (3): 535-57.

Watkins, T., L. Baxter, H. Özkaynak, V. Isakov, and D. Mobley. 2012. Novel approaches for estimating human exposure to air pollutants. *Air Pollution Modeling and its Application XXI*: 741-5.

Yang, G., L. Kong, W. Zhao, X. Wan, Y. Zhai, L. C. Chen, and J. P. Koplan. 2008. Emergence of chronic non-communicable diseases in china. *The Lancet* 372 (9650): 1697-705.

Yang, K., S. Peng, Q. Xu, and Y. Cao. 2007. A study on spatial decision support systems for epidemic disease prevention based on ArcGIS. *GIS for Health and the Environment*: 30-43.

Yi, Q., R. E. Hoskins, E. A. Hillringhouse, S. S. Sorensen, M. W. Oberle, S. S. Fuller, and J. C. Wallace. 2008. Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International Journal of Health Geographics* 7 (1): 29.

Zhang, J., H. Shi, and Y. Zhang. 2009. Self-organizing map methodology and Google maps services for geographical epidemiology mapping. Paper presented at Digital Image Computing: Techniques and Applications, 2009. DICTA'09 229-235.

APPENDIX A: GLOSSARY OF SELECTED TERMS USED IN THIS THESIS

CyberGIS Open Service API (Application Programming Interface): the CyberGIS service API is a set of Web service APIs for CyberGIS application integration and associated computation.

ICD -9: Standard of a set of codes for international statistical classification of diseases and related health problems.

Scalability: The ability of a system, network, or processes for handling a growing amount of work in a capable manner or the ability to be enlarged to accommodate that growth.

Virtual Organization (VO): A dynamic collection of users, resources and services for sharing of cyberinfrastructure resources and services.

APPENDIX B: STRAIGHTFORWARD SPATIAL CLASSIFICATION (SSC) METHOD IN C CODE

main.c

```
#include "hdr.h"

// Parse the command-line parameters that passed by the user
int parsecmdparameters(int argc, char **argv) {
   int i;

   if(argc < 5) {
      printf("[ERROR] Incorrect parameters : %s <dimension> <socEcon filename> <MR
filename> <Output filename>\n",argv[0]);
      exit(1);
   }

   som->argc=argc;
   som->argv=argv;
   som->dimension=atoi(argv[1]);
   som->socEconfilename=argv[2];
   som->arcMRfilename=argv[3];
   som->outputfilename=argv[4];
}


// innerDist comparison for quick sort
int scomp(const void *a, const void *b) {
   if (((arcMR *)a)->innerDist > ((arcMR *)b)->innerDist) return +1;
   if (((arcMR *)a)->innerDist > ((arcMR *)b)->innerDist) return +1;
   return 0;
}


// Calculation
void calculation() {

   int i, j, k, m, d;
   float min, max, range;
   float distance;

   // Normalize 1  divide
   for(k=0; k<som->dimension; k++) {
      d = divider[k];
      if (d == -1) continue;
```

```
    for(i=0; i<countyCount; i++) {
        county[i].points[k] /= county[i].points[d];
    }
}

// Normalize 2  range(0~1)
for(k=0; k<som->dimension; k++) {
    min = FLT_MAX;
    max = FLT_MIN;

    for(i=0; i<countyCount; i++) {
        if (min > county[i].points[k])  min = county[i].points[k];
        if (max < county[i].points[k])  max = county[i].points[k];
    }

    range = max - min;
    for(i=0; i<countyCount; i++) {
        county[i].points[k] /= range;
    }
}

// Weighting factor
for(k=0; k<som->dimension; k++) {
    for(i=0; i<countyCount; i++) {
        county[i].points[k] *= weight[k];
    }
}

// calculation inner distance form 0
for(j=0; j<stationCount; j++) {  // 138
    distance = 0;
    for(k=0; k<som->dimension; k++) {  // 197
        if(use[k])  distance += DISTANCEQ(0.0, station[j].points[k]);
    }
    station[j].innerDist = sqrt(distance);
}

// quick sort by innerDist
qsort(station, stationCount, sizeof(arcMR), scomp);

// set up classID from 1
k = 0;
for(j=0; j<stationCount; j++) {  // 138
    // printf("station[%i].innerDist=%f\n", j ,station[j].innerDist);
    k++;
```

```
      station[j].classID = k;

      int found = 0;
      for(i=0; i<countyCount; i++) {
         if (county[i].countyID == station[j].countyID) {
            county[i].classID    =  station[j].classID;
            county[i].innerDist  =  station[j].innerDist;
            found                =  1;
            break;
         }
      }
      if(!found) {
         printf("County not found in socEcom.  countyID=%d, classID=%d, innerDist=%f,
mr=%f\n",
               station[j].countyID, station[j].classID, station[j].innerDist, station[j].mr);
      }
   }

   // calculation
   for(i=0; i<countyCount; i++) {
      if (county[i].isMR) continue;

      min = FLT_MAX;
      for(j=0; j<stationCount; j++) {
         distance = 0;
         for(k=0; k<som->dimension; k++) {
            if(use[k])  distance += DISTANCEQ(county[i].points[k], station[j].points[k]);
         }
         distance = sqrt(distance);
         if (min > distance) {
            min = distance;  m = j;
         }
      }
      //printf("%i=%i ",i,m);
      county[i].classID    =  station[m].classID;
      county[i].innerDist  =  station[m].innerDist;
      county[i].distance   =  min;
      county[i].mr         =  station[m].mr;
   }
}

int main(int argc,char **argv) {
   int lc;
   int i;

   // Current time of day
```

```
    time_t     timenow;
    struct tm *timets;
    char       timebuf[80];

    // Get the current time
    timenow = time(0);

    // Format and print the time, "yyyy-mm-dd hh:mm:ss zzz"
    timets = localtime(&timenow);
    strftime(timebuf, sizeof(timebuf), "%Y-%m-%d %H:%M:%S %Z", timets);
    //printf("time: SOM Classify start  %s  ---------------------------------\n", timebuf);

    som=&somInfo;

    // Parse the command line parameters
    parsecmdparameters(argc, argv);

    // Read somID.csv
    getsocEconfromfile(som->socEconfilename);

    // Read arcMR.csv
    getarcMRfromfile(som->arcMRfilename);

    // Calculate distence from each county, and get min distance
    calculation();

    // Setup classify.csv
    setupClassify(som->outputfilename);

    // Free allocated memory
    free(station);
    for(i=0; i<countyCount; i++)
    free(county[i].points);
    free(county);

    // Get the current time
    timenow = time(0);

    // Format and print the time, "yyyy-mm-dd hh:mm:ss zzz"
    timets = localtime(&timenow);
    strftime(timebuf, sizeof(timebuf), "%Y-%m-%d %H:%M:%S %Z", timets);

    return 0;
}
```

io.c

```
#include "hdr.h"

// Count the number of lines in f from the start of the file and reset the file pointer
int linecount(FILE *f) {
   int  lc;
   char b, c;
   char line[4000];              // buffer for reading first input data line
   char* dummy = fgets(line, 4000, f);

   if(line[0] == 'C' || line[0] == 'G' || line[0] == '') {
      som->fileHasTitle = 1;     // File has a title
   } else {
      som->fileHasTitle = 0;     // File has no title
      rewind(f);                 // Put the file position to the start
   }

   lc=0;                         // Line count
   while ((c=getc(f)) != EOF) {   // While the character isn't the EOF
      b = c;                     // Save b as a previous char
      if ((c)=='\n')             // If a newline
         lc++;                   // add to line count
   }
   if((b) != '\n')  lc++;        // In case of last char in the input file is not a new line
   rewind(f);                    // Put the file position back to the start

   return lc;
}

// Open a file based on given attributes
FILE *open(char *fn,char *attr) {
   FILE *f;

   f = fopen(fn,attr);
   assert(f!=NULL);

   return f;
}

// Close a file
int close(FILE *f) {
   fclose(f);
}
```

```
// Read a socEcon file and populate a soc
int getsocEconfromfile(char *filename) {
   FILE *f;
   int lc;

   f = open(filename,"r");
   lc = linecount(f);

   // Set size and allocate array
   countyCount = lc;
   county  = (socEcon *) calloc(countyCount, sizeof(socEcon));

   countyCount = readsocEcon(f);

   close(f);
}

// Read socEcon file, populate soc list
int readsocEcon(FILE *f) {
   int  i;
   int  cnt;
   char line[4000];             // buffer for reading first input data line

   rewind(f);                   // Put the file position back to the start
   if(som->fileHasTitle) {
      char* dummy = fgets(line, 4000, f);
   }
   cnt=0;

   int   countyID;
   int   dataID;
   char  *p[som->dimension+2];

   while(fscanf(f, "%s", line) != EOF) {

      p[0] = strtok(line, ",");
      i = 1;
      while((p[i] = strtok(NULL, ","))) {
         if(i > som->dimension+1) {
            printf("Too much columns in socEcon.  line=%d, dimension=%d, variable
count=%d\n%s\n",
                  cnt+1, som->dimension, i+1, line);
            return 1;
         }
         i++;
      }
```

```
if(i != som->dimension+2) {
   printf("Variable conunt error.  line=%d, dimension=%d, variable count=%d\n%s\n",
         cnt+1, som->dimension, i+1, line);
   return 1;
}

if(strcmp(p[0],"Use") == 0) {
   use = (int *) calloc(som->dimension, sizeof(int));
   for(i=0; i<som->dimension; i++) {
      use[i] = atoi(p[i+2]);
   }

   continue;
}

if(strcmp(p[0],"Divider") == 0) {
   divider = (int *) calloc(som->dimension, sizeof(int));
   for(i=0; i<som->dimension; i++) {
      divider[i] = atoi(p[i+2]) - 1;
      if(divider[i] >= som->dimension) {
         printf("divider range error.  dimension=%d, divider[%i]=%d\n",
               som->dimension, i, divider[i]);
         return 1;
      }
   }

   continue;
}

if(strcmp(p[0],"Weighting") == 0) {
   weight = (float *) calloc(som->dimension, sizeof(float));
   for(i=0; i<som->dimension; i++) {
      weight[i] = atof(p[i+2]);
   }

   continue;
}

county[cnt].countyID =  atoi(p[0]);
county[cnt].dataID   =  atoi(p[1]);
county[cnt].isMR     =  0;
county[cnt].classID  =  0;
county[cnt].distance =  0.0;
county[cnt].mr       =  0.0;
county[cnt].points = (float *) calloc(som->dimension, sizeof(float));
for(i=0; i<som->dimension; i++) {
```

```
            county[cnt].points[i] = atof(p[i+2]);
        }

        cnt++;
    }
    assert(cnt+3==countyCount);

    return cnt;
}

// Read a arcMR file and populate a station
int getarcMRfromfile(char *filename) {
    FILE *f;
    int lc;

    f = open(filename,"r");
    lc = linecount(f);

    // Set size and allocate array
    stationCount = lc;
    station  = (arcMR *) calloc(stationCount, sizeof(arcMR));

    stationCount = readarcMRs(f);

    close(f);
}

// Read arcMR file, populate arcmr list
int readarcMRs(FILE *f) {
    int  i;
    int  cnt;
    char line[4000];            // buffer for reading first input data line

    rewind(f);                  // Put the file position back to the start
    if(som->fileHasTitle) {
        char* dummy = fgets(line, 4000, f);
    }
    cnt=0;

    int    countyID;
    int    dataID;
    int    somNodeID;
    int    swcode;
    float  lon;
    float  lat;
    int    year;
```

```
   float  death;
   float  population;
   float  mr;

   while(fscanf(f, "%d,%f\n",
            &countyID, &mr) != EOF) {

      //if(year == 0) continue;

      station[cnt].countyID  =  countyID;
      station[cnt].mr        =  mr;

      int found = 0;
      for(i=0; i<countyCount; i++) {
         if (county[i].countyID == countyID) {
            station[cnt].points  =  county[i].points;
            county[i].isMR         =  1;
            county[i].classID      =  0;
            county[i].mr           =  mr;
            found              =  1;
            break;
         }
      }
      if(!found) {
         printf("County not found in socEcom.  line=%i, countyID=%d, mr=%f\n",
              cnt+1, countyID, mr);
      }

      cnt++;
   }
   return cnt;
}


// Save the Classify.csv
int setupClassify(char *filename) {
   FILE *f;
   f = open(filename,"w");
   writeArcMR(f);
   close(f);
}

// Write an ascii grid file
int writeArcMR(FILE *f) {
   int i;
```

```
   for(i=0; i<countyCount; i++) {
      fprintf(f, "%d,%d,%d,%f,%f\n",
            county[i].countyID, county[i].isMR, county[i].classID,
            county[i].distance, county[i].mr);
   }
}

// Write socEcon array for demo
int writesocEcon(char *filename) {
   int  i;
   FILE *f;
   f = open(filename,"w");
   fprintf(f,
"CountyID,dataID,isMR,classID,innerDist,distance,MR,popHden,popDen,popDenM,popDenF,n
onAg\n");

   for(i=0; i<countyCount; i++) {
      fprintf(f, "%d, %d, %d, %d, %f, %f, %f, %f, %f, %f, %f, %f\n",
            county[i].countyID, county[i].dataID,    county[i].isMR,
            county[i].classID,  county[i].innerDist, county[i].distance, county[i].mr,

county[i].points[0],county[i].points[1],county[i].points[2],county[i].points[3],county[i].points[4])
;
   }
   close(f);
}

// Write arcMR array for demo
int writearcMR(char *filename) {
   int  i;
   FILE *f;
   f = open(filename,"w");
   fprintf(f, "CountyID,classID,innerDist,MR,popHden,popDen,popDenM,popDenF,nonAg\n");

   for(i=0; i<stationCount; i++) {
      fprintf(f, "%d, %d, %f, %f, %f, %f, %f, %f, %f\n",
        station[i].countyID, station[i].classID, station[i].innerDist, county[i].mr,

station[i].points[0],station[i].points[1],station[i].points[2],station[i].points[3],station[i].points[4]);
   }
   close(f);
}

// Display to console
int displayArcMR() {
   int i;
```

```
    for(i=0; i<countyCount; i++) {
        printf("%d,%d,%d,%f,%f\n",
                county[i].countyID, county[i].isMR, county[i].classID,
                county[i].distance, county[i].mr);
    }
}
```

hdr.h

```
#include <stdlib.h>
#include <stdio.h>
#include <assert.h>
#include <string.h>
#include <float.h>
#include <math.h>
#include <time.h>
#ifdef _OPENMP
#include <omp.h>
#endif

// Data structures

// A socEcon.csv
typedef struct {
   int    countyID;          // county ID  input
   int    dataID;            // data ID  input
   int    isMR;              // is MR from arcMR.csv  output
   int    classID;           // class ID  output
   float  innerDist;         // distance from 0,  when isMR == 1
   float  distance;          // distance  output
   float  mr;                // mr  output
   float  *points;           // socioeconomic and  demographic variables
} socEcon;

// A arcMR.csv
typedef struct {
   int    countyID;
   int    classID;
   float  innerDist;          // distance from 0
   float  mr;
   float  *points;            // socioeconomic and  demographic variables
                              // Point to socEcon's points.

} arcMR;

// Store important SOM information for the program
typedef struct {
   int argc;
   char **argv;
   int  fileHasTitle;
   int  dimension;            // dimension of the variables in the socEcon file
   char *socEconfilename;     // socEcon filename
   char *arcMRfilename;       // arcMR filename
```

```
   char *outputfilename;          // output filename
} somstr;

// Definitions of static variables
#define PRINTPOINTS 0
#define PRINTGRID   0
#define INT_MAX     2147483647
// #ifndef M_PI
// #define M_PI 3.14159265358979323846
// #endif

// Global variables
int     countyCount;
int     stationCount;
socEcon *county;
arcMR   *station;
somstr  somInfo, *som;
int     *use;                  // set 1 if the variable is used to calculate the result of SSC
                               // set 0 if the variable is not used to calculate the result of SSC
int     *divider;              // Divider column;  -1 means no dividing
float   *weight;               // Weights

#ifdef _OPENMP
#else
long countcell;
long countall;
long countselected;
#endif

// Macros
#define max(a,b) a>b?a:b
#define min(a,b) a<b?a:b
#define DISTANCEQ(p1,p2) ((p1)-(p2))*((p1)-(p2))

// Function definitions
FILE *open(char *fn,char *attr);
int linecount(FILE *f);
```