CLUSTER ANALYSIS OF HEALTH DATA INVOLVING GENERAL DISEASE CLUSTER SHAPES AND MULTIPLE VARIABLES

DISSERTATION

Presented to the Graduate Council of
Texas State University-San Marcos
in Partial Fulfillment
of the Requirements

for the Degree

Doctor of PHILOSOPHY

by

Zhijun Yao, M.S.

San Marcos, Texas

May 2011

CLUSTER ANALYSIS OF HEALTH DATA INVOLVING GENERAL DISEASE CLUSTER SHAPES AND MULTIPLE VARIABLES

Committee Members Approved:

F. Benjamin Zhan, Chair

Yongmei Lu

Nate Currit

Ram Shanmugam

Approved:

J. Michael Willoughby
Dean of the Graduate College

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

ACKNOWLEDGEMENTS

My foremost thanks go to my dissertation advisor, Dr. F. Benjamin Zhan. This dissertation could not have been possible without his excellent guidance, support and patience during my graduate study. He inspired and advised me by questioning my assumptions, discussing my hypothesis, improving my methodology, improving and polishing my dissertation. Without his excellent guidance and constant support and encouragement, this dissertation could not have been finished. I also would like to express my sincere gratitude and appreciation to my outstanding committee members, Dr. Yongmei Lu, Dr. Nate Currit, and Dr. Ram Shanmugam for their thoughtful advice and mentoring.

I would like to give my special thanks to the Texas State University-San Marcos faculty and staff. Very special thanks to Dr. Lawrence Estaville for being an encouraging supporter and showing selfless concern for students during my graduate study.

I would like to give my deep gratitude to my family and friends who were always supportive for me. Especially, I thank my parents for their endless love and encouragement throughout my schooling. I would like to thank my wife, Junmei Tang, and my sons, Alex and Brad, who give me such support during my PhD studies. Thank all the friends I have at Texas State. Thank everyone who helped me in this dissertation.

This Manuscript was submitted on December 12, 2010.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

CLUSTER ANALYSIS OF HEALTH DATA INVOLVING GENERAL DISEASE

CLUSTER SHAPES AND MULTIPLE VARIABLES

by

Zhijun Yao, B.S., M.S.

Texas State University-San Marcos

May 2010

SUPERVISING PROFESSOR: F. BENJAMIN ZHAN

This dissertation research develops new methods to automatically detect clusters of general shapes in disease data involving multiple variables. Disease cluster detection is important in public health surveillance as well as in disease control and prevention. A number of techniques were proposed and developed for identifying compact clusters, yet few of them can detect clusters with irregular shapes efficiently. Furthermore, most researchers failed to notice the combined influence of multiple variables on human disease, focusing only on the impact resulting from a single variable. This dissertation research has two primary objectives. The first one is to develop two new methods, a maximum-likelihood-first algorithm and a non-greedy algorithm, which can be used to detect disease clusters of arbitrary shapes. These new methods were applied to detect clusters of murine typhus disease cases in southern Texas from 1996 to 2006. The second objective is to develop a procedure for detecting line segment clusters based on visual exploration (parallel coordinates technique) and spatial analysis techniques. Similar to, but unlike the parallel coordinates technique, this procedure can be used to detect concentrations of the simultaneous occurrence of the instances of two properties represented by two variables.

Although numerous projects have focused on cluster analysis of disease point data, this research is the first systematic investigation on both point and line clusters. This dissertation research contributes to the literature of spatial cluster analysis in detecting arbitrary-shaped patterns of disease point data. The method and results from the line cluster interpretation helps public health officials utilize line cluster techniques in identifying the relationship of multiple variables with the help of visualization techniques.

CHAPTER 1

INTRODUCTION

1.1. Background

In recent years, there has been a dramatic increase in the public awareness of environmental hazards and disease events (Yiannakoulias et al. 2007). More and more people are concerned with the report of increasing incidence of a disease taking place in their or nearby communities, especially when there is some environment exposure linked with the incidence of the disease. Each year, public health departments in the United States receive more than one thousand requests from the public for disease cluster investigation, though a majority of them are ultimately found to be false alarms (Greenberg and Wartenberg 1991; Trumbo 2000).

Public health officials have to act cautiously in response to these requests. For individuals who have a family member or friend suffering from a disease, this is a tragedy and they feel it is necessary to find out the disease source. For the general public, requesting dismiss without any scientific reason is considered unacceptable even though public health professionals are dubious about the existence of disease cluster for public (Wartenberg 2001). Based on the experiences from previous studies (Alexander 1999), most requests which asked for further fieldwork investigation were not practical and had to be declined. Furthermore, given limited resources, it is impossible for public health departments to embark on an investigation for every request brought by a concerned public or interested media. A better strategy is to discover underlying causes that leads to anomalous disease patterns and to explain them to the public. Sometimes disease clusters occur by chance alone, and sometimes they are caused by genetic factors or habits like smoking or drinking. However, in many cases, these clusters arise from disease outbreaks which can be caused by unknown infectious agents, pollution sources, environment hazards, or adverse health risk conditions. For example, evidence of an aggregation of disease occurrences around a pollution source may suggest that a further investigation is necessary for the elevated risk of developing a particular disease.

Cluster research has played an important role in modern epidemical research and public health practice. It can also play a role as pre-epidemiology (Wartenberg and Greenberg 1993). In order to efficiently and effectively deal with the reports of disease clusters, CDC (Centers for Disease Control and Prevention) published the “Guidelines for investigating clusters of health events” in 1990 (CDC 1990). Most, if not all, researchers utilize exploratory tools to help them generate epidemiological hypotheses for cluster analysis. A proper specific hypothesis might lead to further studies to acquire the etiological and pathological knowledge to identify the causes of diseases. As Rothman pointed out that “the payoff from clustering research comes from the specific hypotheses that emerge to explain the observed pattern of excess occurrence” (Rothman 1990).

Detection methods are also critical in disease cluster analyses. Given the wide adoption of statistical methods in geographical research since the early 1960s (Burton 1963), it is not surprising that there has been considerable development of relevant statistical techniques in geographical epidemiology. In particular, medical geographers as well as statisticians and epidemiologists have developed a large number of statistical tests to facilitate disease cluster analysis and investigation. For example, Kulldorff (2006) reported more than 100 tests in the literature. There have also been several reviews of frequently used statistic methods (Besag and Newell 1991; Marshall 1991; Moore and Carpenter 1999; Bailey 2001; Pickle 2002; Sankoh and Becher 2002; Kulldorff 2006).

Among numerous statistical test methods, the scan statistic method, also known as moving window analysis, has received increasing attention recently with respect to cluster analysis. Scan statistics was first introduced by Naus (1965) aiming at detecting anomalous patterns in one-dimension data. The principle idea is to search over a great many windows (intervals) superimposed over the data and carry out some statistical test based on a specific data distribution for each window. The windows with relatively high test statistic are of interest.

The first application of scan statistics in spatial cluster detection was the GAM (namely Geographical Analysis Machine) developed by Openshaw et al. (1987), but the spatial scan statistic method proposed by Kulldorff has been most widely used (Kulldorff and Nagarwalla 1995; Kulldorff 1997). The Kulldorff's spatial scan statistic is a relatively simple but powerful method for the detection of spatial clusters. It searches over a number of circles which are centered at each case location or the centroids of each component area, using various radii. Then, it calculates a test statistic for each window with a pre-specified likelihood function. In the calculation, the scan window which has the maximum statistic is selected as the most likely cluster and its statistical significance is evaluated by Monte Carlo simulations. More details about the spatial scan statistic will be provided in chapter 3.

Compared to other spatial cluster analysis methods, Kulldorff’s technique has several advantages. The first one is its simplicity. It has no parameter needed to be determined in advance, that is, it doesn't require that users have a prior knowledge with respect to any potential clusters. The second one is that it has the capacity to identify the most likely cluster and shows its location with an evaluation using the statistical significant values at the same time. The last one allows for uneven underlying population densities using scan windows in variable sizes, and adjusts for the multiple hypothesis testing problem using Monte Carlo simulation. This method is put into practice in the free software SatScan which can be downloaded from the website http://www.satscan.org/. The method, along with the software, is widely used today in many applications, especially in public health surveillance and disease prevention and control. Kulldorff (1999) provides a long list of applications of spatial scan statistic across a wide range of fields.

1.2. Study Objectives and Research Questions

The overall objective of this dissertation research is to develop methods that can be used in cluster analysis of health data involving general disease cluster shapes and multiple variables. Specifically, this work focuses on two tasks. One is to develop new statistical methods capable of detecting disease spatial clusters which may be of unknown shapes. The other one is to apply the cluster analysis method in detecting the relationship of multiple variables, with the help of visualization techniques.

Technically, this research intends to explore a new cluster detection method and to compare its performance with several established methods, aiming to improve the capability and accuracy of automated cluster pattern detection techniques. This research is guided by the following questions:

(1) How do cluster tests detect clusters of general shapes rapidly and automatically? For this research, the murine typhus disease has been selected as a case study to identify its cluster pattern.

The application of cluster detection in public health research is primarily used to detect the unexpected aggregation of disease outbreaks and provide timely information for disease surveillance. These outbreaks may be caused by environmental hazards or bioterrorist attacks. To examine the question above, this study investigates the performance of automatic cluster techniques in detecting “anomalous” spatial patterns which might be an indication of an emerging epidemic. A new cluster method is developed and compared to the commonly-used scan statistic method.

(2) Which scan statistic technique is more effective to detect the ‘true’ spatial pattern of disease data – the one with regular shape windows or the one with irregular shape windows?

The spatial scan statistic, one of the most effective statistic techniques used in cluster pattern detection, has been extended from one dimension data description to two or three-dimension data exploration. In practice, the spatial scan statistic method adopts the variable windows with different shapes and sizes. Most researchers commonly employ scan windows with regular shapes, such as circles (Openshaw et al. 1987; Turnbull et al. 1990; Kulldorff and Nagarwalla 1995; Hjalmars et al. 1996; Alm 1997; Kulldorff 1997), ellipses (Kulldorff et al. 2006), rectangles (Alm 1997; Anderson and Titterington 1997) and triangles (Alm 1997). However, the outbreak of disease does not always follow regular shapes and this will hinder the subsequent investigation of the diseases in question. To answer this question, this study investigates the effectiveness of irregular window shapes in disease cluster pattern analyses.

(3) How to detect the impact of simultaneous occurrence of multiple variables, and is it possible to use the line cluster test to find out the potential combined impact from multiple variables? Is visual exploration technique efficient to detect the clustering of line segments?

Most cluster detection methods so far have concentrated on the analyses related to spatial point patterns. These point cluster analyses are derived only from the distances and distributions between each pair of disease incidences without considering the impact from nearby environmental factors. Based on the previous research findings from the point data, this study explores the opportunities of visual exploration and spatial analysis techniques in detecting the concentration of the occurrences of two variables.

1.3. Contributions and Significance

This dissertation research presents a significant contribution in the literature of cluster analysis of spatial data. Building upon the existing cluster detection techniques, this project adopts alternative tactics to construct a set of scan windows with arbitrary shapes including circular windows. They are expected to identify clusters with more complex structures than those detected using the traditional circular spatial scan statistic. The result of line cluster detection will contribute to the body of knowledge regarding how multiple environment variables co-impact to the identified disease and the distance where the highest co-impact exists from multiple environmental factors. The new statistical techniques and results have the capacity to provide timely and practical information to help manage the spread of a disease. The methods can be useful for scientific and educational purposes.

The primary significance of this research can be appreciated from both technical and application perspectives:

(1). From a technical perspective, this study develops an advanced algorithm to detect disease clusters automatically, especially for clusters with arbitrary shapes. By extending this scan statistic algorithm to line networks, the application domains of cluster detections has been extended from point patterns to line patterns. This new algorithm will also make significant contributions in other applications where the detection of clusters of line segments is important.

(2). More significantly, this study provides a valuable tool to health departments and residents regarding disease cluster analysis. Using the proposed automatic cluster detection method, this research result will provide the tool capable of generating scientific evidence about potential outbreaks of disease which could then be used to warn the public at local, state, or even national levels. Moreover, it demonstrates that the arbitrary-shape cluster detection can identify emerging epidemics with higher spatial accuracy than the traditional scan statistic method.

1.4. Structure of the Dissertation

The remainder of the dissertation is divided into six chapters. Chapter 2 focuses on the review and discussion of related literature, and Chapter 3 describes the study sites and procedures about data preparation. Chapters 4 presents the theoretical and statistical framework to summarize Kulldorff's spatial scan statistic, the basic principle this dissertation build upon. Chapter 5 develops two new algorithms that can be used to detect spatial clusters with arbitrary shapes, using the murine typhus incidences in southern Texas as a case study. Chapter 6 combines visual exploration and scan statistic method to analyze linear clusters and discusses its potential applications to examine the combined impact of two variables. Chapter 7 discusses the research results, potential applications, and future research directions.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews previous research about cluster detection and related statistical methods. The review will start from the definition of cluster and cluster analysis. Then, the importance and applications of cluster analysis, particularly in epidemiology, will be discussed. Finally, various statistical methods employed in cluster analysis will be reviewed.

2.1. Disease Cluster and Cluster Analysis

Theoretically, clusters can be defined as an unusual aggregation of events which are represented as a high concentration of events or values in time and space (Sun 2008; CDC 2009). From a geographer's perspective, cases inside these spatial regions are closer to each other than cases outside. The basic goal of spatial cluster detection is to identify the places where the observed incidents excess the expected incidents significantly in statistical terms, adjusting for underlying inhomogeneous population or other covariates such as age or gender. A cluster may provide useful clues for an emerging/existed disease outbreak.

The spatial aggregation or cluster analysis has been used by researchers in a variety of disciplines, ranging from biological studies of DNA (Leung, Choi, and Chen 2005) to environmental studies of national pollution (Diggle, Rowlingson, and Su 2005). For example, cluster analysis methods could be applied to unusual patterns in DNA sequence to find biological origins of diseases (Leung, Choi, and Chen 2005). Similarly, in an environmental study, it might be necessary to identify anomaly of distributions in space and/or time domain that have higher density than expected. Generally, the application of cluster detection aims to find out the anomalous pattern and tendency of clustering. That is, cluster analysis will try to find out whether this unexpected pattern exists and where it happens.

Clearly, it is impossible to interpret the result of clusters appropriately without defining the scale of measurements (Marshall 1991). This scale can be hundreds of miles to several miles over space or tens of years to several weeks over time. Comparing to a larger scale, an apparent spatial cluster within a small homogeneous area is more worthwhile and attracts more public attention since it is easy to identify the potential mechanism for the cluster (Rothman 1990). For example, one essential application of cluster analysis is to detect unexpected distribution of disease cases within the observed area (Neill 2006). The research of cluster detection, particular in the public health and epidemiology, has been explored by a variety of statistic models and quantitative methods.

2.2. Cluster Detection in Public Health and Epidemiology

2.2.1. Definition of a Disease Cluster

There still remains a controversy about the definition of the term “disease cluster”. Many definitions of disease cluster have been proposed in terms of magnitudes of excessiveness of incidences (Caldwell and Heath 1976; Cook-Mozaffari et al. 1989; CDC 1990; Heath 1996; Moore and Carpenter 1999; Wartenberg 2001). Some of them are rather general, for example, a disease cluster can be defined as “an unusual aggregation, real or perceived, of health events that are grouped together in time and space and that are reported to a health agency” (CDC 1990) or “an aggregation of cases in an identifiable subpopulation” (Wartenberg 2001), while some others are quite specific, such as “five cases representing at least a five-fold increase in risk have to be seen by a single physician (or a small group of close colleagues) over a short time” (Lancet 1990) or at least five cases with a high relative risk (Neutra 1990). These studies, taken together, suggest that the definition of disease cluster depends on the specific disease.

A more qualitative definition, provided by Knox (1989), defines disease cluster as “being a geographically and or temporarily bounded group of occurrences of a disease already known to occur characteristically in clusters, or of sufficient size and concentration to be unlikely to have occurred by chance, or related to each other through some social or biological mechanism, or having a common relationship with some other event or circumstance”.

None of these definitions, however, discusses how to tell whether the aggregation is a true cluster or not. According to CDC statistics, less than 5% reported clusters can be confirmed to be “true” clusters after investigations. Many researchers expressed their concerns over definition issues involved in disease cluster investigations (Rothman 1990; Wartenberg 1995, 2001; Wakefield, Quinn, and Raab 2001; Bachmann 2003). In these studies, the disease cluster was defined as a region where a statistically significant increase of disease incidence are observed than expected in a given area during a specified time period among a particular population group.

2.2.2. Early Work in Cluster Exploration of Public Health Data

Disease Mapping

The earliest statistical exploration of disease cluster in epidemiology and public health could be traced back to the early nineteenth century. An early well-known example is Dr. John Snow's ingenious work on cholera epidemic (Snow 1855). During the formidable epidemic of cholera in 1854 in London, he used a map plotting the distribution of water pumps and victims of the fatal disease in the city. The map revealed there was a readily apparent aggregation of incidents around a public water pump on Broad Street (Figure 1). The abnormal aggregation led him to posit that the pestiferous source of cholera was the contaminated water instead of breathing foul air that the public were concerned with. A further investigation confirmed his hypothesis. His discovery of the means of spreading cholera brought the epidemic to an end after blocking the contaminated water pump. His research started an era of disease mapping in epidemiology.

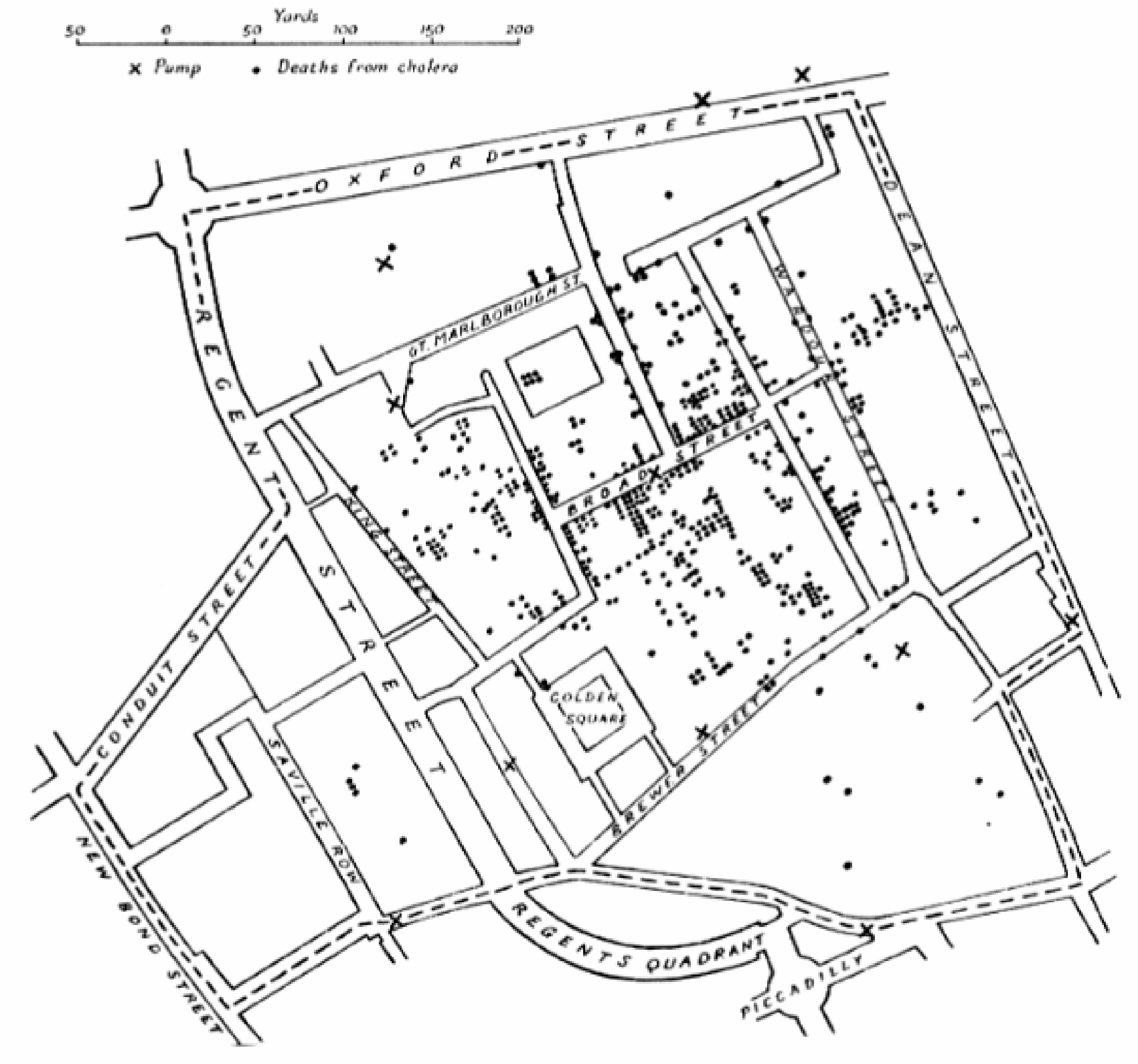


Figure 1. The original John Snow’s famous cholera epidemic map (Source: Snow 1855).

Followed the similar method as John Snow, the Hamburg cholera epidemic in 1892 attracted the public attention to disease mapping again (Lawson and Williams 2003). Associated with other social economic data, such as population, the dot maps showed the case of cholera in the cities of Hamburg and Altona and indicated the potential risk source. Holden (1880) mapped the mortality and sanitary in New Jersey and recognized that the absence of sewage systems was the major reason of typhoid. The maps of typhus during 1922 to 1925 in Montgomery, Alabama further confirmed the formation of a rodent-borne disease. The value of these early disease maps was demonstrated their usefulness in helping determine the aetiology of an infectious disease.

There are various ways to represent disease on maps, using formats such as proportional rates, grey scale, dot density, contour, or 3-Dimention plots. Using GIS (Geographical Information System) tools, many research projects were funded to produce small-scale disease distribution maps (Andes and Davis 1995; Bayers et al. 1996; Popovich and Tatham 1997; Hightower et al. 1998). Most early disease mappings tried to identify the suspected areas at high disease risk and some maps may help to verify the hypotheses about disease aggregation.

Generally, early disease mapping played an important descriptive role in spatial epidemiology, however early mapping methods only obtained a “good” visual estimation of geographic distribution of the disease over the study area (Bailey 2001). Not surprisingly, when more disease cases with precise locations are available, there is an increasing demand on the application of quantitative or statistical method on pattern recognition and significance assessment. More advanced spatial statistics are needed to do a further spatial analysis.

Early Practices in Disease Clustering

Disease clustering always occurs if there is some direct or indirect of risk factors, such as biological (Blum 1948), environmental or social (Fraser et al. 1977; Baptiste et al. 1984). It is necessary to track the pattern or process of these outbreaks to identify potential risk factors. More and more statistical techniques are proposed to model and represent disease incident distribution and process. Clayton and Kaldor (1987) developed an empirical Bayesian statistic to estimate the age-standardized risk. This statistic was further explored by Clayton and Bernardineli (1992) to map lip cancer in Scotland and breast cancer mortality in Sardinia. Cliff and Haggett (1988) used join count statistics to reveal a clustering pattern of cholera cases in London. Douven and Scholten (1995) summarized the statistic measurements on disease point data into three ways: based on distance, density, and inter-point distance. Kitron and Kazmierczak (1997) used Moran's I to explore the pattern of Lyme disease in Wisconsin between 1991 and 1994.

There are tremendous statistical methods applied to disease cluster analysis and these applications depend on the development of general statistical methods. In the following section, we will review the major categories of spatial cluster analysis method with a focus on two automatic cluster detection methods: the geographical analysis machine and the spatial scan statistic.

2.3. Spatial Cluster Analyses and Major Statistical Methods

2.3.1. Traditional Spatial Statistics

Early spatial pattern studies were mostly descriptive using visual or mathematical methods, such as mean center or standard distance (Rogerson and Yamada 2009). The *descriptive statistics* is a typical nonspatial statistics, aiming at summarizing the characteristics of spatial data distribution. The typical descriptive statistics include mean center, median center, range, percentiles, variance, deviation, standard distance, skewness, and kurtosis. Descriptive statistics measure the fundamental geographic concepts such as location, dispersion, and moments (David 2005). Specifically, the location analysis could be performed by Mean, Median or Mode while the dispersion could be analyzed by Range, Standard Deviation, Percentile, or Coefficient of Variation. The moments, which are other import geographic concepts, could be represented by Skewness, Kurtosis or Variance, and Semivariance (David 2005).

Distinguished from descriptive statistics, *inferential statistics* aim to support inferential statements to a dataset or make comparisons between sets of data (David 2005). The typical inferential statistical research studies include the construction of confidence intervals, statistical estimation, and hypothesis testing (Rogerson and Yamada 2009). Although much emphasis has been placed on these traditional statistic models which provide foundation for spatial clustering analysis, most of them provide just one simple statistic to represent a single variable or trend and this highly limits their further applications to more complex data.

Besides the distance, another commonly used test function for clustering is the density of points. For instance, *kernel density estimation,* one technique used to estimate density, is non-parametric through averaging the observed data by a known kernel function (Elgammal et al. 2002). Typically, an intensity surface is generated by estimating the intensity of the grid points covering the whole study area (Rogerson and Yamada 2009). Since no particular kernel function is required and any density function could be selected, this technique is applicable to many clustering analyses when the underlying density is unknown (Scott 1992; Duda et al. 2000).

Similar to kernel density estimation, *k-function* is another popular statistical measurement which aims to examine second-order characteristics of point data (Wong and Lee 2005; Rogerson and Yamada 2009). First-order characteristic and second-order characteristics differ in that the former measures the mean of a process over space while the latter emphasizes on the spatial process resulting from spatial dependency (Rogerson and Yamada 2009). K-function partitions the observed data into k clusters and compares it with the expected value (Ripley 1981; Rogerson and Yamada 2009). K-function is now widely used in analyzing spatial pattern of vegetation (Peter 1995), bird nests (Gaines et al. 2000), soil microbes (Nunan et al. 2002), traffic accidents (Jones et al. 1996), and disease incidents (Diggle and Chetwynd 1991).

2.3.2. Categories of Statistical Methods for Cluster Analyses and Related Problems

Before embarking on a review of various spatial cluster analysis methods, it is useful to group them into different categories. In terms of cluster dimensions, the tests could be grouped into spatial, temporal, and spatial-temporal cluster analysis. Given the data we use, these tests could be classified into point pattern analysis and area pattern analysis. They can also be grouped into distance, nearest neighbors, and autocorrelations according to the relationships that are measured.

A more commonly used classification is provided by Besag and Newell (1991) based on the purpose of these statistics: test of clustering and test for the detection of clusters. Test of clustering could be further subdivided into two subcategories: global (general) statistics and focused statistics. Global statistics attempt to use a single test statistic to assess overall clustering tendency of disease incidence over the whole study area (Jacquez 2008). It can indicate whether disease incidence clusters or not. Various random processes may result in global clustering (Haining 1998). One example is the spread of infectious disease where the disease is transmitted among nearest neighbors. Another example is that a large number of environment hazards such as lead-painted houses over the region of interest, which causes the excess of disease incidences at many locations. Focused statistics are usually used when there is a specific predefined hypothesis which concerns possible links between the aggregation of disease incidences and suspected risk factors.

However, both the global and focused statistics have shortages as we apply them to clustering analysis. Specifically, global statistics can neither pinpoint the specific locations of potential clusters nor provide additional information on the sizes and shapes of clusters if they do exist. Focused statistics only show interests in a few particular locations where suspicious environmental exposures may account for clusters. These locations could be point pollution sources such as power plants or waste disposal sites, or linear sources like power lines or highways. A hypothesis cannot be generated from the distribution of data themselves; otherwise it will lead to the Texas sharp shooter problem (the Texas sharp shooter fires his gun at the wall of a barn first, and then draw the bulls-eye around the bullet holes to show how he is good at shooting). Kulldorff (1998) called such methods “evaluating cluster alarms” when we observe a local excess from data distribution first and then want to assess its statistical significance.

The second type of tests, test of the detection of clusters (also called local statistics), is concerned with finding local clusters. These tests have two goals: 1) identifying locations, sizes and shapes of potential clusters, and 2) assessing whether the detected clusters are statistically significant or occur by chance. Usually, test of the detections of clusters is of more practical than test of clustering because it suggests the locations of significant disease clusters for further investigations.

Generally speaking, each statistic category has its strengths and weakness in some particular applications. Before the cluster analyses are performed, each category should be well-known to select the method most appropriate for cluster analyses. In the following section, each detailed category will be reviewed.

2.3.3. Global Statistics

Theoretically, most early spatial cluster analyses are global by providing a single statistical summary without identifying exact cluster sizes and locations (Jacquez 2008). In the early global statistical research, the observed values of global statistics were compared with the expected statistical values upon which to accept or reject the null hypothesis (Rogerson and Yamada 2009).

Originally developed in the field of ecology, the *quadrat method* and the *nearest neighbor statistic* are two early methods to test whether the geographic distribution of species is spatially random or not (Rogerson and Yamada 2009). The quadrat method was initially developed to explore the characteristics of point distribution patterns (Haggett, Cliff, and Frey 1965; Diggle 2003). Typically, the quadrat method counts the points that fall in the quadrants which are divided from the study area. The parameters generated in this process, including the location of each quadrat, the number of point falling inside each quadrat, the adjacency among these quadrats, will be incorporated to perform the spatial analyses.

The nearest neighbor method, originally proposed by J. G. Skellam (1952), is based on the ratio between the mean value of the nearest neighbor distance and the expected value of the nearest neighbor distance. The ratio less than one indicates a clustering tendency and the relatively dispersed pattern could be found with the ratio larger than one. This method was further explored by Clark and Evans (1954) through incorporating a statistical test of significance. Although the quadrat method and the nearest neighbor statistic method have been proven successful in many applications, they are quite limited when applied to the disease cluster analysis since the disease distribution is complicated which varies with many factors such as age, occupation, gender, income, environments etc.

Developed by Moran (1948), *Moran'I* is by far the most widely used measurement of spatial autocorrelation. The larger absolute value of Moran's I represents a higher autocorrelation among the observations and zero values represented random spatial distribution. The values smaller than zero represent negative relationships while the values larger than zero represent positive relationships. Moran's I has been applied to hundreds of applications for spatial clustering since it was published, including methodology estimation (Cliff and Ord, 1972), population density (Assuncao and Reis 1999), geographical data properties (Bennett and Haining 1985), public health (Walter 1992), and species distribution (Carl and Kuhn 2007).

Many other statistic indices based on the Moran's I, have been developed to detect global clustering. These statistics include *Geary's C statistic* (1954), *Grimson's* *method* (1989), *Cuzick-Edwards Test* (1990), *Besag and Newell's method* (1991), *Getic and Ord's global statistic* (1992), *Oden's I_{pop} statistic* (1995), *Tango's statistic* (1995), and *spatial Chi-Square statistic* (1997). The null hypothesis for these global statistical methods is usually defined as "no clustering exists". Testing this null hypothesis based on one statistical value, global statistics do not provide a significant assessment for any particular locations which might lead to an error in missing the significant local "spots" of incidents (Rogerson and Yamada 2009).

2.3.4. Focused Statistics

Different from global statistics, focused *statistics* are only applied to some specific locations (Jacquez 2008). The majority of these statistics are practically appropriate to detect possible clusters near the source of environmental problems such as the disease cluster centers around the pollution source (Puett et al. 2005; Puett et al. 2009). For example, Waller et al. (1992) applied the focused statistics to identify leukemia clusters near the groundwater sites which were affected by hazardous trichloro ethylene in upstate New York. Lawson and Williams (1994) examined the spatial clusters of respiratory cancer in Armadale from 1968 to 1974 using focused statistics. In their paper, they also suggested including the non-parametric kernel regression to detect the population at risk from the putative pollution sources.

Environmental exposure is still the major interested application of the focused cluster test. Along with this interest, various cluster detection methods have been proposed and applied. Bithell (1995) proposed incorporating the spatial functions such as inverse distance into the focused-cluster since the impact from the pollution will decline when the distance to the source increases. This method was further improved in Bithell's article (1999) for the disease mapping using the relative function. Through plotting the relative risk function (RRF), Bithell's method counts a risk score for each case and summarizes these risk scores to derive the test statistic. This method is also called linear risk score (LRS) test due to its linear structure.

Diggle (1990) proposed an inhomogeneous Poisson process model to evaluate the pattern of Laryngeal cancer next to a disused industrial incinerator in Lancashire, England. Using the same dataset, Diggle and Rowlingson (1994) reanalyzed the impact of pollution from three industrial plants on the disease by a modified conditional approach. These focused statistics researches investigate the possible linkage between the increased risk of larynx and lung cancer and the suspicious environment factors.

Developed by Lawson (1989) and Waller et al. (1992), the score test is another popular focused statistical method. This method tests the frequency of spatial pattern around some particular point-focus under the hypothesis as "no clustering around the focus". In this method, the inverse of distance to the focus is used to estimate the strength of environmental pollution. By accumulating the exposure strength from nearby pollution sources, each region got a score and this score is used to test whether the relationships among the environmental factors and the focused sites exist or not. More effects, such as peaked effect, direction effect, are combined using the mathematical functions in the Lawson (1993) research on the mortality events using pre-defined points.

2.3.5. Local Statistics

Local statistics, with many of them derived from global statistics, quantify spatial clustering within the small areas and these small areas could cover the entire study area (Jacquez 2008). The most famous system of local statistics is the *Local Indicators of Spatial Association* (LISA) (Anselin 1995) which is a set of local statistics decomposed from global Moran'I statistics.

Many other local statistics were developed upon Anselin's LISA method: Tango (1995) developed a modified score statistic *Tango's C_F Statistic* to test clusters around perspecified locations; *Besag and Newell* (1991)'s local statistic version is developed to screen the clusters for childhood leukemia; *Getis' G_i Statistic* (Ord and Getis 1995) was developed from global statistic to measure the local clustering tendency.

More complex statistics are developed when the spatial trend, such as population heterogeneity, is to be modeled. Clayton and Kaldaor (1987) brought the spatial Gaussian prior into the likelihood estimation for relatively common diseases in disease mapping. Clayton and Bernardinelli (1992) incorporated prior knowledge into cancer mapping. Besag et al. (1991) used the Gibbs sampler for counts and over dispersion in image restoration. Turnbull et al. (1990) detected local spatial clusters of leukemia cases. Kulldorff (1997) developed the famous scan statistic to identify the unusual pattern of cases in space and time. The detailed information about the spatial scan statistic will be reviewed in the following section.

However, one disadvantage of these local statistical methods is their huge computational task. They have to search over an enormous number of regions to test the multiple hypotheses and this is not practical for massive real-world datasets. If less restrictive constraints are applied or a large number of hypotheses need to be tested, this situation will be even worse (Openshaw et al. 1988; Neill 2006). Automatic cluster detection, a new method to analyze the data without any preconceived hypothesis, becomes very popular in cluster analysis.

2.4. Automatic Cluster Detection

2.4.1. The Geographical Analysis Machine (GAM) and its Development

Developed by Openshaw and his colleagues in 1987 and 1988, the GAM (Geographical Analysis Machine) is an early effort to look for spatial patterns in an automatic manner. It has been commented as “the first major attempt to identify clusters of a rare disease”(Besag and Newell 1991, p.148). The GAM employs an exploratory approach to indicate the possible locations of clusters so that a hypothesis might be derived. The approach is carried out as follows: GAM lays out a grid mesh over the study area of interest. Then a large number of circles with a fixed radius are generated as possible clusters (Figure 2). They are centered at each grid point location and the radius is chosen to be a little bit larger than the grid spacing so that neighboring circles have a certain degree of overlap. GAM repeats the procedure with a range of radii so that all possible clusters with different sizes in the study area are guaranteed to be found. These circles are examined in an exhaustive way in order to find those in which disease incidence exceeds the expectation derived from a Poisson distribution. Their statistical significances are then evaluated by Monte Carlo simulations. Those circles with elevated incidence but lower significance than a given statistics level are drawn on the map. Those areas with most intensity of circles are of interest for further investigation.

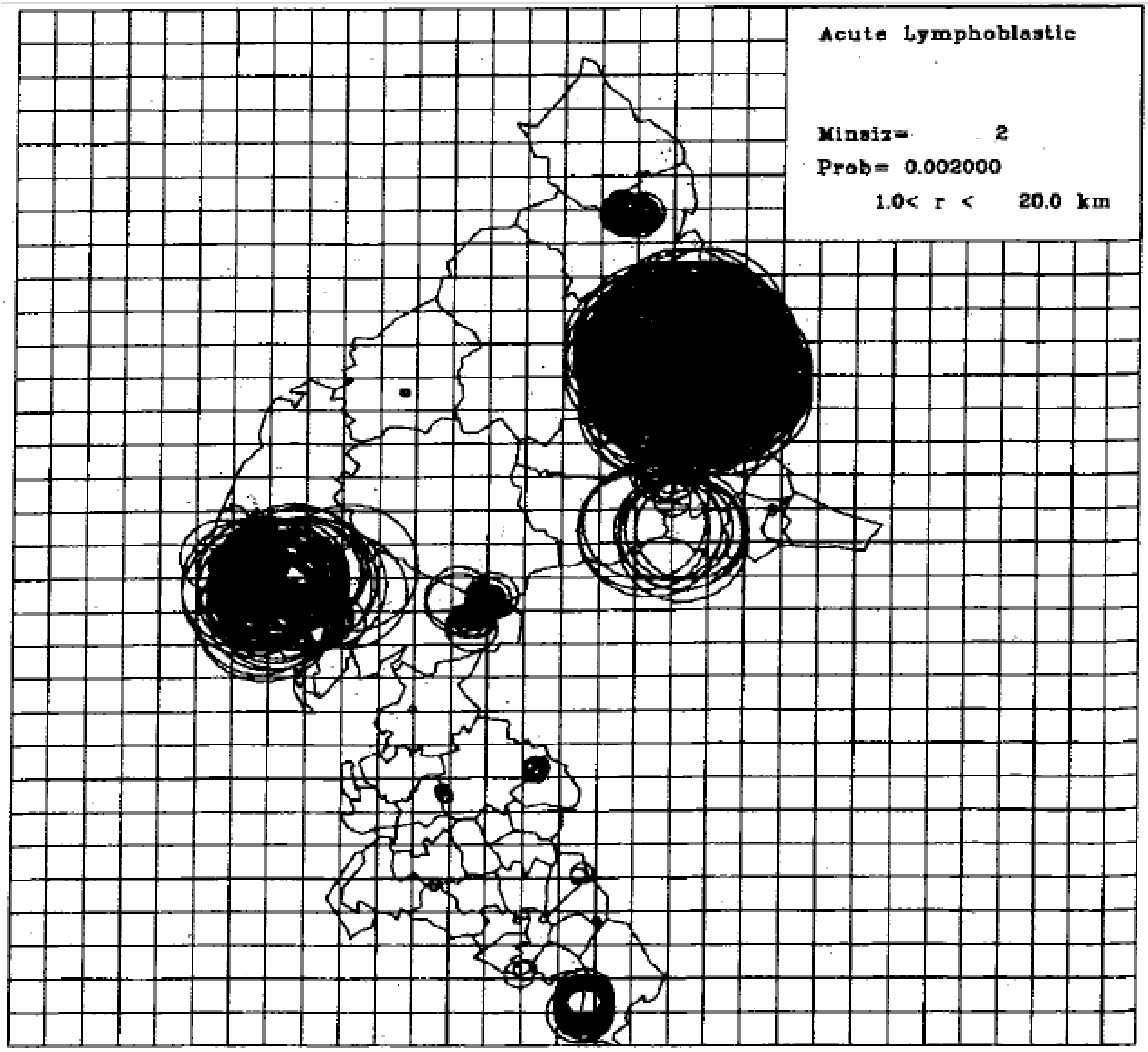


Figure 2. Openshaw's GAM applied to childhood leukaemia in England (Source: Openshaw 1987).

The GAM attracts some criticisms. First, it does not allow for risk covariates such as sex and age. Second, it searches over a large number of circles in an exhaustive way requiring enormous computational workload. Third, it is hard to calculate the number of incidences and population susceptible to the disease in the generated circular areas because the real data are usually aggregated into administrative districts which have irregular shapes (Besag and Newell 1991). It is also heavily criticized because its ad hoc statistical basis and the massive generated overlapping circles lead to the problem of multiple testing, that is, many false positive clusters are bound to be detected (Besag and Newell 1991; Marshall 1991; Kulldorff and Nagarwalla 1995). The GAM carries out statistical tests for each area separately to examine whether the area is a false cluster given a fixed significant level (α). For example, $\alpha = 0.05$, corresponds to a cluster can be falsely detected with the probability 0.05. That is, if 10,000 circles are tested separately, there would be 500 expected false clusters. Moreover, if an area is tested 10,000 times independently, at least one false cluster would be detected because the probability will be $1- (1-0.05)^{10000} \approx 1$. Marshall (1991) pointed out that this multiple testing problem may distort the discovery of true clusters. Despite these drawbacks, GAM is a successful method because it avoids the problem of preselection bias, uses data at a fine resolution and avoids urban-rural bias resulting from either fixed-size or fixed population subregions (Besag and Newell 1991). The GAM has inspired the development of several related methods.

Fotheringham and Zhan (1996) presented a “refined” GAM in attempting to alleviate the computation workload by examining fewer circles. They proposed generating circles to be tested in a random manner rather than exhaustively. Instead of systematically generating numerous circles at each grid point, the proposed method creates circles whose locations are randomly assigned and whose sizes are arbitrarily chosen from a specified range. A sufficient number of circles would be generated and tested so that there is little chance of "left out" for any part of the study area during the entire procedure. Compared with GAM, fewer circles are examined; therefore, it significantly eases the computational workload which is heavily criticized in GAM. In addition, since the method examines fewer circles, the result correspondingly contains fewer false clusters. Regarding to the significance testing, they adopted the Poisson probability of observed counts within a circle as their test statistic instead of using observed counts directly.

Rushton and Lolonis (1996) developed a method to search for spatial clusters of birth defect for the people living in the urban. This method is similar in spirit to Openshaw's GAM (1987) except three slight differences. First, Rushton and Lolonis chose radii smaller than the grid spacing, therefore, fewer circles would share some particular incidents. Second, they adopted incident rate instead of incident count as the test statistic in this method. This modification takes into account the underlying heterogeneous population distribution. And third, a different strategy was applied in Monte Carlo simulations. The expected incidences were generated according to the overall risk rate and the probability obtained from a uniform distribution, resulting in the total number of cases varies in each simulation.

The Cluster Evaluation Permutation Procedure (CEPP), introduced by Turnbull et al. (1990), is another method modified upon the GAM (Openshaw et al. 1987). In the previous research (Openshaw et al. 1987; Fotheringham and Zhan 1996; Rushton and Lolonis 1996), the collections of circles are generated with a geographic or Euclidean distance as their radius. So even for the same radius, circles would contain inhomogeneous population at risk and various observed incidences. Turnbull et al. (1990) proposed to use circles at various geography sizes. Within each circles, he predefined the number of population who are at risk. Therefore, the collection of circles would contain a constant population size but vary in geographic size since the population density is heterogeneous. The predefined population size is achieved in such a way that a circle around a region will absorb the population from its nearest neighbors. When the number of persons residing in the nearest regions is aggregated, there is little chance to get a circle with the exact population specified in advance. Therefore, it would be better to add a portion of population from the last nearest region to reach the required population, as well as that corresponding portion of cases. As the collection of circles has constant population, the comparison is only needed to make comparisons with the amount of observed cases within each circle. Instead of getting a lot of clusters as reported in other methods, the CEPP is only interested in the place where the most likely cluster is located. So in this method, the test statistic value is selected from the any circle which has the maximum case number, referring to that of the most likely cluster. Since it is sensitive for the selection of population size, Turnbull et al. (1990) suggested running their test with different population sizes in order to get an optimal result.

In their test, Turnbull et al. (1990) introduced an innovative way to test the significance level by adopting the Monte Carlo method to make the simulated data. Unlike other tests, their significance test only deals with a single test statistic and a single reference distribution. For one particular circle, the previous tests compare the test statistic between the real data and the simulated data. Instead, their test makes comparisons between the most likely cluster observed and the most likely clusters from the simulations, which could be any circle at any place in each simulation. Then, the statistical significance is evaluated as the probability that the ratio of the most likely clusters from the real data is no less than those from the simulations. They employed this strategy in order to avoid or at least reduce the problem of multiple testing. However, in their method, they required the cluster detection test to be applied with different population sizes in order to get an optimal result. This requirement inevitably introduces the multiple testing problem again, because it does not provide the comparison across circles with various size of population.

Besag and Newell (1991) used another way to improve GAM. In contrast to the CEPP, in which Turnbull et al. (1990) used the fixed number of population as the radii to construct circles, Besag and Newell (1991) concentrated their attention on circles which have a constant case number for a rare disease. In other words, they explicitly determined the size of cluster first, and then tried to search for those circles which have the most likely clusters among a collection of circles with the pre-specified number of cases. This method was implemented as following: first put a circle at the centroid of each area which contains non-zero cases and then include its nearest neighboring areas in the order of the ascending distances until the collection of areas contain at least pre-specified number of cases. Unlike the CEPP, this method does not require adding a portion of cases or population to obtain the difference from the included neighbor. The required number of areas to construct such a circle is used as their test statistic. A small number indicates a likely cluster. The method uses Monte Carlo method to evaluate the significance of detected likely clusters. The simulated cases for each area are generated under the null hypothesis that these areas have a constant risk rate and each area has an independent Poisson distribution. They examine the probability that the prescribed number of cases will be observed in a circle made up by fewer areas than that observed from the real data. Besag and Newell (1991) suggested plotting these circles with smaller significances than a given level on a map as the presence of the most likely clusters. Furthermore, Besag and Newell (1991) also suggested to use the global clustering index which is the total number of detected clusters that fall below a given significance threshold. Since this presented method searches for clusters in a selective manner, it is able to reduce the processing time compared with the GAM (Openshaw et al. 1987). This method also bears some limitations. One is that its significance evaluation procedure inevitably results in multiple testing problem and this problem is further exaggerated when testing for a range of cluster sizes. The other problem is it requires a fixed size of cluster which is subjectively determined by users. Thus this method limits the power to detect clusters with variable sizes. If a small size is given, the method would stop when it achieves the required minimum number of study cases, leaving out some possible large clusters to be assessed or just returning subclusters of large clusters; if a large size is given, and then it would miss many small clusters.

The results of the preceding tests are sensitive to somewhat unknown parameters which should be set before they start the program. For instance, the CEPP (Turnbull et al. 1990) needs to generate circular windows with an unknown constant population size. In Besag and Newell's (1991) research, they set the cluster size to an unknown number. While the models developed by Openshaw et al. (1987), Rushton and Lolonis (1996), and Fotheringham and Zhan (1996) required a predefined distance radius. When we attempt to apply these tests, we have to figure out appropriate values for these parameters in advance, which are to some extent arbitrary and hard to be determined by the non-professional users. But unfortunately, the choice of these parameter values is likely to affect the test result. Therefore, we have to repeat these procedures with a range of parameter settings. Moreover, all of these proceeding tests encounter the problem of multiple testing. A new method is needed to solve these problems.

2.4.2. A Detailed Review on Spatial Scan Statistic and its Derivations

Inspired by the work of Openshaw et al. (1987) and Turnbull et al. (1990), Kulldorff (1997) developed the famous spatial scan statistic and it can be used to detect clusters of various sizes as well as account for the multiple testing problem. Kulldorff (1997) treated the CEPP model (Turnbull et al. 1990) and the model developed by Rushton and Lolonis (1996) as two special cases of his general approach. By placing a circular window on the map, the spatial scan statistic will move across the study area as shown in the Figure 3. The position of the centroid can be the area centers, the case locations, or other different coordinates. Rather than specifying the size of a potential cluster a priori, this method uses a scan window of varying sizes, corresponding to varying population and varying number of incidents. The radius of windows increases

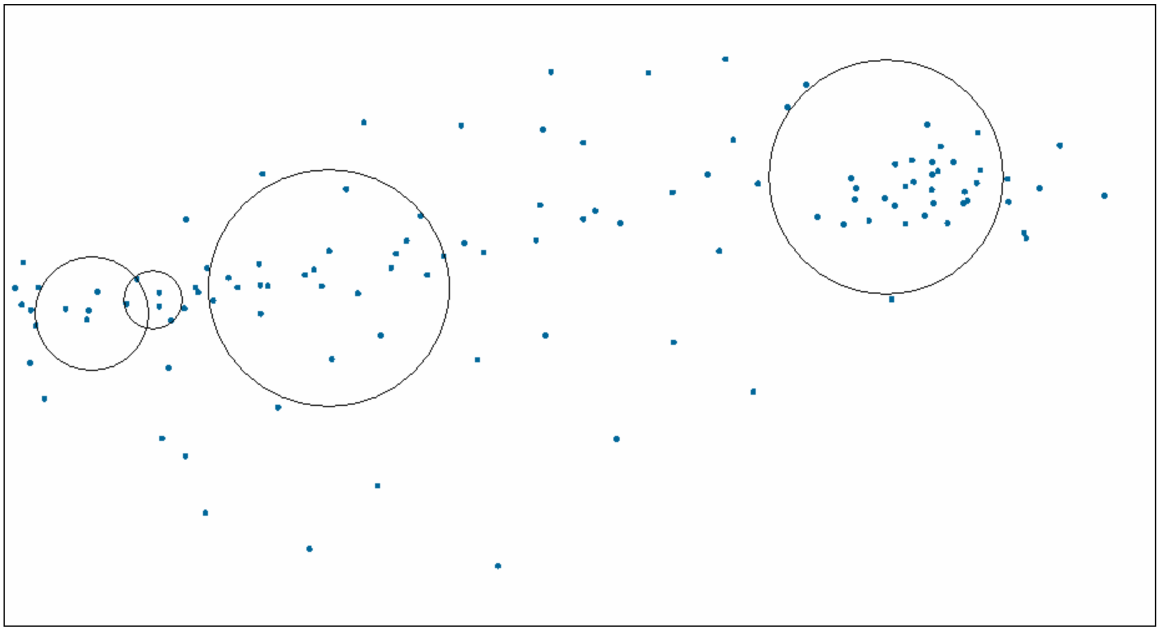


Figure 3. A subset of circular scan windows.

continuously from zero till an upper radius which is predefined based on either a population percentage or a geographical size. And then the method will calculate a likelihood ratio for each window (as a circle) according to the probability of the observed and expected number from Poisson model or Bernoulli model within this window. The window with the highest value is identified out to be the most likely cluster. Once the most likely cluster is detected, its statistical significance is evaluated by Monte Carlo simulation. Rather to examine the significance of a particular circular window at a given level, the spatial scan statistic evaluates the probability of getting a more extreme likelihood ratio from the simulations than the value of the most likely cluster from the real data. Since these maximum likelihood ratios are acquired independently between the simulations, it provides a valid adjustment for the problem of multiple testing.

In addition to the most likely cluster, Kulldorff (1997) suggested using the same way to search for secondary clusters which have lower likelihood ratios. They are of less use if little additional information would be provided. Since adding or removing a fewer areas to or from the most likely cluster exerts little influence on the values of likelihood, the likelihood ratio of the secondary clusters and the most likely cluster would be almost the same if the secondary clusters overlap with the subset areas of the most likely cluster. Thus those secondary clusters that are nonoverlapping with the most likely one should be highlighted as our interesting clusters. The significance of these secondary clusters is assessed through comparing their likelihood values with the maximum likelihood distribution obtained from simulations. The reason to do so is that the null hypothesis should be rejected, if the secondary cluster is a real one, no matter the most likely cluster is true or not. This also implies that no secondary cluster is significant enough if the most likely cluster is not.

Generally speaking, the spatial scan statistic has the following features (Kulldorff 1999): 1) it takes account of inhomogeneous population density and other confounding variables, for example, gender or age; 2) it avoids the problem of pre-selection bias by searching clusters without any a priori assumption on the size and location; 3) it corrects the inherent problem of multiple testing when many possible sizes and geographical locations of clusters are taken into consideration; 4) it indicates the approximate location of a cluster when it results in the rejection of the null hypothesis; 5) it is able to detect clusters with higher risks as well as clusters with lower risks. Furthermore, a recent performance comparison shows that this method is the most powerful test in detecting a compact cluster (Kulldorff, Tango, and Park 2003). Due to these advantages, the spatial scan statistic has rapidly become very popular and it is widely used in a large number of applications.

Derivations Based on Scan Windows with Fixed Shapes

Based on the circular spatial scan statistic, Kulldorff (1999) proposed an extension of his test two years later, which was named as the isotonic spatial scan statistic. This new method takes account of the relationship between risk and distance, that is, it makes an assumption that the risk is higher within a certain distance from a cluster center than the risk beyond that distance. Thus the risk can be modeled using a step function in which the risk decreases with increasing distance to a cluster center.

Kulldorff (1999) suggested using an isotonic regression function to obtain the highest likelihood.

Kulldorff (2001) also extended the spatial scan statistic for both spatial and temporal clusters. In the proposed space-time scan statistic, a three-dimensional cylindrical window in various sizes is used. Similar as the spatial scan statistic method, the base of the cylinder is a circular window to represent an exact particular geographical area and use the cylinder height to represent one of the time intervals of the study period. The cylinder is flexible in the location, circular geographical base size, its starting date and time interval. The cylinder is moved in space and time so that a large number of overlapping clusters varying in size and height are obtained, examined and evaluated. The cylinder with the highest likelihood represents the most likely cluster.

Later on, Kulldorff et al. (2006) provided an elliptical spatial scan statistic. Unlike the circular spatial scan statistic which uses circles as scanning windows, the new method attempts to use ellipses to define cluster areas. An ellipse is controlled by location, shape, orientation, and size. The locations are represented by the centroids' x coordinates and y coordinates which are usually corresponding to the centroids of areas. The shape of the moving ellipse could be defined by the ratio between the length of semimajor axis (the longest axis) and semiminor axis (the shortest axis). To reduce the computational workload, a specific number of shapes are chosen such as 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 15, 20, 30, 60 and 120 (Kulldorff et al. 2006). An ellipse would be degraded to a circle when the shape equals to one, that is, the semimajor axis and the semiminor axis have the same length. The ellipses' orientation could be defined by the angle between the semimajor axis and the axis in the horizontal direction. The angles are chosen so that there is an around 70 percent overlap among the ellipses regarding to their shape, location or size. The size of an ellipse ranges from zero to a pre-specified upper limit such as half of the total population. In order to reduce the number of eccentric ellipses, which are of less interest in the analysis, Kulldorff et al. (2006) introduced an eccentricity penalty function to adjust the likelihood ratio. Their results show that the elliptical spatial scan statistic has a good performance in detecting clusters of circular or elliptical shapes. Finally, Kulldorff et al. (2006) emphasized that the location and size is critical for detecting clusters, whereas the choice of the shape of the scanning window, irrespective of circle or ellipse, is of less importance because the most likely cluster merely provides an indication of the general area of a true cluster, whose exact boundary usually is vague.

Based on Kulldorff's model, Neill and Moore introduced a fast scan algorithm (Neill and Moore 2004; Neill 2006). The primary objective is to speed up the performance of the spatial scan statistic when a large dataset is under consideration and the second objective is to detect elongated clusters. The algorithm is based on a novel overlap-kd tree data structure dividing the rectangular study area into overlapping subregions, and a top-down search approach. It searches large subregions first and then their smaller subregions. It bounds regions including their subregions which can obtain the maximum likelihood ratio, and prunes regions and their subregions that are not the most likely cluster. This method is likely to only examine a subset of subregions, consequently considerably reducing the search time. Since it uses rectangular regions as scan windows, it is able to detect axis-aligned elongated clusters.

Derivations Based on Scan Windows with Arbitrary Shapes

A major critique to Kulldorff's method is its predefined-shape scan windows which prevent its application from the arbitrary cluster patterns. As mentioned above, it uses windows in circles or other fixed shapes with various sizes to detect potential clusters, though Kulldorff (1997) pointed out in his paper that they could be any other shapes. Using circles or rectangles or any other predefined geometrical shapes as scan windows restricts the patterns of disease clusters to be detected and leaves a large number of candidate clusters out of the test. Empirical results show that spatial scan statistic performs well in identifying compact clusters, but poor in dealing with elongated or arbitrary-shaped clusters (Kulldorff, Tango, and Park 2003). However, disease clusters in the real world could appear in any shapes. For example, pathogens, one kind of disease source, may disperse in a lot of ways. The wind can carry airborne pathogens in certain directions. The running water may spread waterborne pathogens along the path of a river or a stream. Similarly, pathogens may disperse along a road or other transportation routes. Under these conditions, the disease incidence patterns will shape as elongated clusters. Therefore, the shapes of disease clusters are potentially dependent on the way how disease incidences occur and disperse. Furthermore, even the inhomogeneous underlying population at risk will distort the shape of clusters. In addition, Tango and Takahashi reported that the spatial scan statistic based on circular windows is likely to include neighboring areas without elevated risk which result in a larger cluster detected than the true one (Tango 2000; Tango and Takahashi 2005). How to detect arbitrary-shaped clusters presents a challenge for researchers.

Recently, many researchers attempted to solve this problem and proposed many solutions. These new methods adopted the identical statistical principles behind the circular spatial scan statistic but differentiate in the collection of candidate scanning windows. They introduced different strategies for the construction of scanning windows of irregular shapes. For example, Patil and Taillie (2004) introduced the concept "upper level set", a set of areas having a larger risk rate than a given constant, to detect arbitrarily shaped hotspots. Duczmal and Assuncao (2004) introduced the famous "simulated annealing strategy", one of the global optimization methods, to search the local maxima for the arbitrarily-shaped spatial clusters detection.

More searching strategies are proposed based on some advanced algorithms. Assuncao et al. (2006) brought forward a more generalization strategy, minimum spanning tree, to reduce the number of neighbors to be searched. A genetic algorithm is employed to limit the irregular shape of clusters in order to find the most potential real clusters (Duczmal et al. 2007). Wieland et al. (2007) introduced Euclidean minimum spanning trees to locate any noncircular clusters. Yiannkoulias et al. (2007) presented two approaches to improve the greedy growth search: one is the non-connectivity penalty to limit the very irregular cluster shapes and another one is the depth limit to prevent the generation of large super-clusters from smaller clusters.

One common feature of these derivations to detect the arbitrary shape clusters is their consideration of adjacency characteristics of areas instead of using predetermined geometry shapes. They search for irregular-shaped clusters among adjacent areas based on the assumption that any subset of adjacent areas could make up a potential cluster and the shape of this cluster might not be circular or rectangular. Since exhaustive search might be implementation infeasible, various constrains are set to guide the search process so as to reduce the number of candidate scan windows.

With these various derivations developed, the spatial scan statistic becomes popular, even though it still has some drawbacks which restrict its applications. The spatial scan statistic has less capability in dealing with a large dataset (Neill 2006) which makes its approach insufficient when applied to the large dataset such as large-scale health surveillance. In addition, spatial scan statistic has a limited number of statistical models available for underlying data (Duczmal and Assuncao 2004; Patil and Taillie 2004; Tango and Takahashi 2005; Neill 2006). Kulldorff (1997) gave two special discrete models in his paper, namely, Poisson model and Bernoulli model. He limits its application to some real phenomena since continuous or other type data are very common in many other cases. For example, Kulldorff's approach restricts its application on network spaces such as river networks or highway systems (Patil and Taillie 2004). If these drawbacks can be overcome, the spatial scan statistic will certainly play a critical role in the spatial cluster analysis, not only for the public health, but also for other applications.

CHAPTER 3

THEORETICAL AND STATISTICAL FRAMEWORK OF SPATIAL SCAN STATISTIC

Before presenting the study method and algorithm, this chapter will introduce the theoretical and statistical framework to summarize Kulldorff's spatial scan statistic. Kulldorff's spatial scan statistic consists of at least three parts: choosing a distribution model of data, identifying the most significant cluster, and testing its statistical significance.

3.1 Choosing a Distribution Model of Data

Spatial scan statistic carries out the likelihood ratio test based on two particular models: Bernoulli distribution and Poisson distribution. Both models are applied to calculate the probability of observed phenomena. In a Bernoulli model, people are described by zeros or ones, correspondingly representing cases or controls. These cases and controls make up the total population. In a Poisson model, the number of people with a disease in an area follows a Poisson distribution. In other words, the expected cases in an area are proportional to its population at risk.

Generally, spatial scan statistic attempts to identify clusters with a risk rate which is statistically significantly high. It makes an assumption that the observed cases are drawn from a chosen distribution model. The null hypothesis (H_0), that is, no cluster, against a set of alternative hypotheses (H_1) that there is a cluster in a given area will be tested. If the null hypothesis is true, the risk rate for any area should be the same as the risk rate of the entire study area. More precisely, it is defined as:

H_0: the underlying risk rate is a constant for all areas.

$H_{1:}$ the underlying risk rate of an area is higher than those rates outside the area.

The basic theory of spatial scan statistic is to test the likelihood ratio for the study area. It compares the likelihood of observed data to the likelihood which derived from the null hypothesis. For the sake of simplicity, the same notations and equations will be adopted as Kulldorff's (1997) in the following part. Let **Z** be the set of regions generated by the scan circles, **μ(G)** represent the total population of the study area, **nG** denote the total observed case number, **μ(z)** stand for the population of **z**th region**,** and $\mathbf{n_z}$ be the observed number in **z**th region. Meanwhile, ***p*** is defined as the probability that an incident falls in the **z**th region, ***q*** as the probability that an incident falls in the rest of the study area, and ***p***, ***q*** are the numbers between 0 and 1. Now a specific region z could be tested whether it is a cluster. The null hypothesis is it has a constant probability for all areas (H_0: ***p***=***q***, **z** є **Z**) and the alternative hypothesis is the specific area z has a larger ***p*** than ***q***s of the outside areas (H_1: ***p***>***q***, **z** є **Z.)**.

For a given region **z**, the likelihood function based on the Bernoulli model can be expressed as the following formula:

$$L(z) = \sup_{p>q} L(z,p,q) = (p)^{n_z} \times (1-p)^{\mu(z)-n_z} \times q^{nG-n_z} \times (1-q)^{(\mu(G)-\mu(z)-(nG-n_z))} \qquad (1)$$

The observed likelihood function on region **z** can be given by

$$L(z)=\begin{cases}\sup\limits_{p>q} L(z,p,q)=(\dfrac{n_z}{\mu(z)})^{n_z}\times(1-\dfrac{n_z}{\mu(z)})^{\mu(z)-n_z}\times & \\ (\dfrac{nG-n_z}{\mu(G)-\mu(z)})^{nG-n_z}\times(1-\dfrac{nG-n_z}{\mu(G)-\mu(z)})^{(\mu\mu(G-\mu(z)-(nG-n_z))} & if\ (\dfrac{n_z}{\mu(z)})>(\dfrac{nG-n_z}{\mu(G)-\mu(z)}) \\ & \\ or & \\ & \\ (\dfrac{nG}{\mu(G)})^{nG}\times(\dfrac{\mu(G)-nG}{\mu(G)})^{\mu(G)-nG} & otherwise\end{cases}$$

(2)

Thus the expected likelihood function can be derived using:

$$L_0=\sup_{p=q} L(Z,p,q)=(\frac{nG}{\mu(G)})^{nG}\times(\frac{\mu(G)-nG}{\mu(G)})^{\mu(G)-nG} \qquad (3)$$

Therefore the likelihood ratio λ(z) can be obtained as the quotient through dividing the observed likelihood by expected likelihood:

$$\lambda(z)=\begin{cases}\dfrac{L(z)}{L_0}=\dfrac{\sup\limits_{p>.q} L(z,p,q)}{\sup\limits_{p=q} L(Z,p,q)} & if\ (\dfrac{n_z}{\mu(z)})>(\dfrac{nG-n_z}{\mu(G)-\mu(z)}) \\ & \\ or & \\ 1 & otherwise\end{cases} \qquad (4)$$

Kulldorff (1997) also calculate the likelihood ratio test based on the Poisson model as following:

$$\lambda(z)=\begin{cases}\dfrac{L(z)}{L_0}=\dfrac{(\frac{n_z}{\mu(z)})^{n_z}\times(\frac{nG-n_z}{\mu(G)-\mu(z)})^{nG-n_z}}{(\frac{nG}{\mu(G)})^{nG_z}} & if\ (\dfrac{n_z}{\mu(z)})>(\dfrac{nG-n_z}{\mu(G)-\mu(z)})\\ \\ or\quad 1 & otherwise\end{cases}\qquad(5)$$

3.2. Identifying the Most Significant Cluster

Given the above likelihood ratio formulas based on either one of the two models, $\lambda(z)$ can be easily calculated for a set of possible regions. The test is carried out through the following steps. To begin with, it gets the centre positions of the component areas of the map. Then it will associate the number of disease incidents and the population of a corresponding area to these positions. In the second step, a moving circle window is placed on the study area and position of the centroid of the circular window can be the centre of census position or other different coordinates. The size of the moving window changes from zero to pre-selected value. Each circular window defines a region which is made up by a set of areas whose centroids reside in the circle. And then a likelihood ratio for each region will be calculated by comparing the observed likelihood and expected likelihood according to a distribution model. Finally, the test will sort the likelihood results within each circle and choose the maximum result as the most likely cluster.

Since no prior knowledge about possible clusters is available when running these tests, they are usually used “blindly” which might result in two types of inevitable errors in cluster detections (Neill 2006). The first type of error, false positives, occurs when the insignificant clusters are mistakenly treated as true clusters. While the second type of error, false negatives, treats the true clusters as false ones. These errors must be controlled during our tests because both of them may result in high costs for public health. The first type error may mislead health departments to spend precious limited resources and efforts on unnecessary investigations while the second type error may delay the interventions which could control and prevent an emerging disease outbreak effectively. To enhance the capability of detecting the true clusters, these errors should be minimized. That is why the third step, significance assessment, is essential.

3.3. Testing Statistical Significance

Once the most likely cluster has been identified, the next step will be the test that whether "this potential cluster" occurs due to a disease break or just by chance. To do so, *p*-value, derived from the Monte Carlo simulation, is used to assess the statistical significance for the detected cluster.

The Monte Carlo simulation was proposed by Dwass (1957) and was first introduced to cluster detection tests by Turnbull et al. (1990). In a Monte Carlo simulation, a large number of random replications will be generated under a chosen distribution model conditioned on the simulated case number will be the same as the real data. If the Poisson model is adopted to test the null hypothesis, then the real population from each area will be used in this replication. The disease events in each area are drawn from an inhomogeneous Poisson distribution with mean $\mu(z)\frac{nG}{\mu(G)}$, where $\mu(z)$ denotes the population of area z, $\mu(G)$ represents the total population of the whole study area and *nG* is the total observed number. The likelihood ratio for each region was calculated using the replica data as well as the real data and plenty simulations were performed. Each simulated data will get a maximum likelihood ratio in the same way as the real data. Then *p*-value can be calculated based on the sorted likelihood ratio of the real data and simulated data. For example, if there are *N* simulated datasets and one real dataset and the total number of datasets will be *N+1*. Within these total datasets, there is *n* simulations having a larger or equal maximum likelihood ratio compared to the one obtained from the real data. That is, the rank of the real data is *n* when we sort the data by their maximum likelihood ratios. The *p*-value for the significant testing in this example will equal to *n*/(*N+1)*. Theoretically, the smaller the p-value, the more significant the cluster will be.

CHAPTER 4

DATA PREPARATION AND PROCESSING

In this chapter a detailed description of data preparation and processing will be provided. Since this dissertation was conducted from two research focuses, cluster detection of general disease cluster shapes and multiple variables analysis, two datasets were collected for each aspect. Dataset one is the murine typhus cases reported in the South Texas and second one is Texas births data and public toxic substances data from U. S. Department of health and Human Services.

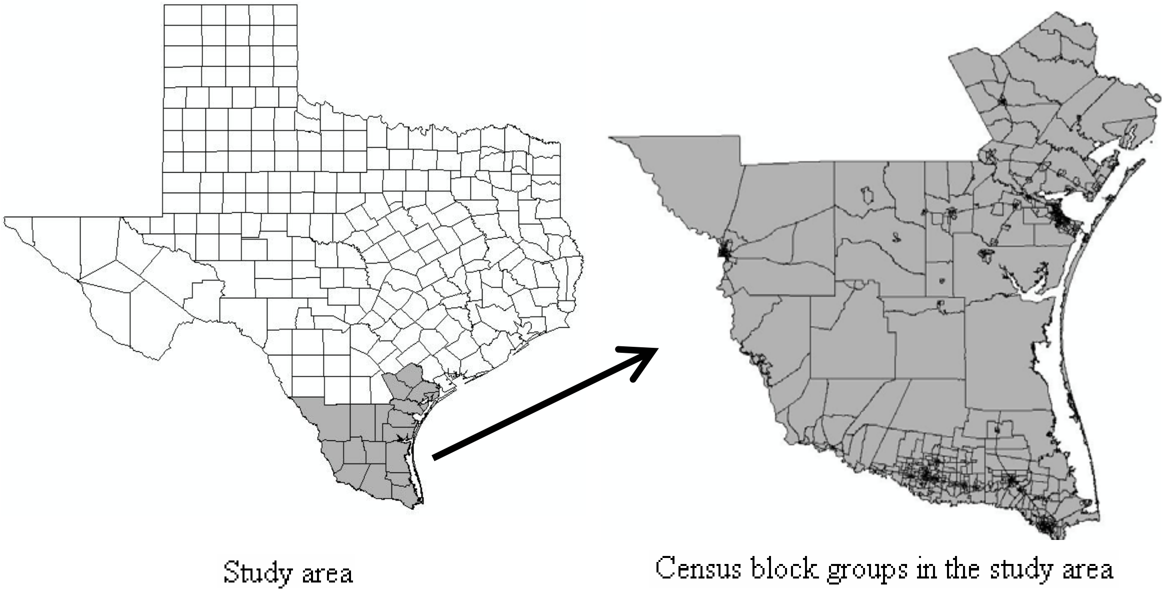
4.1. Data to Support Detection of General Disease Cluster Shapes

4.1.1. Data Preparation

The study area for our first research is south Texas, one of the areas having the most murine typhus cases occurred in United States. From the 1970s, the murine typhus cases were reported around 20 cases/year in this area (Boostrom et al. 2002). Centered at 98° 18' W longitude and 27° 12' N latitude, the study area includes 17 counties of in south Texas with population around 2 million.

The data used in this research include census block group boundary data, population data, and disease data issued by the department of health. In this research, the cluster detection was performed at the census block group level in our cluster detection (Figure 4). The data were obtained from ESRI website and DVD (ESRI 2008). The entire study area covers 1,068 census block groups with 1,728,393 inhabitants (U. S. Census

Bureau 2000). The disease data used in this research consist of 555 murine typhus case records reported to the Health Department of Texas in the south Texas from 1996 to 2006. The raw disease data are stored in an Excel file, containing the geographical location of cases (latitude and longitude), the onset time of cases (year, month, and day), age, gender, and race of patients, zip code and street name of cases. Although these cases are reported throughout the whole year, 44% of cases are found in May, June, and July.


Study area

Census block groups in the study area

Figure 4. Study area and census block groups used as study boundary.

4.1.2. Data Processing

The original census data and disease data come in various data types and formats. The census block group file was downloaded as a GIS shape file, while population and disease data were acquired in a text file format. In order to analyze and display all the data in the GIS platform, it is imperative to join the population data and disease data into the same format as the boundary file. The excel tables of population data were converted into MS Access tables and joined to boundary files at census block group level. The demographic information is essential in the following research since they provide the number to calculate the probability for each unit.

The disease data were spatial joined to the census block groups using software ArcGIS 9.3. The disease data are provided in Excel file with two accuracy levels: zip code and street. When joined disease data to the census block groups, only the records at the street accuracy level were used and the records at the zip code level were excluded. Thus, after the data processing, there were 555 murine typhus cases at the zip code level and 391 cases for the analysis at the census block group levels. The distribution of murine typhus is illustrated in Figure 5, 6, and 7 at zip code area level, census tract level, and census block group level.

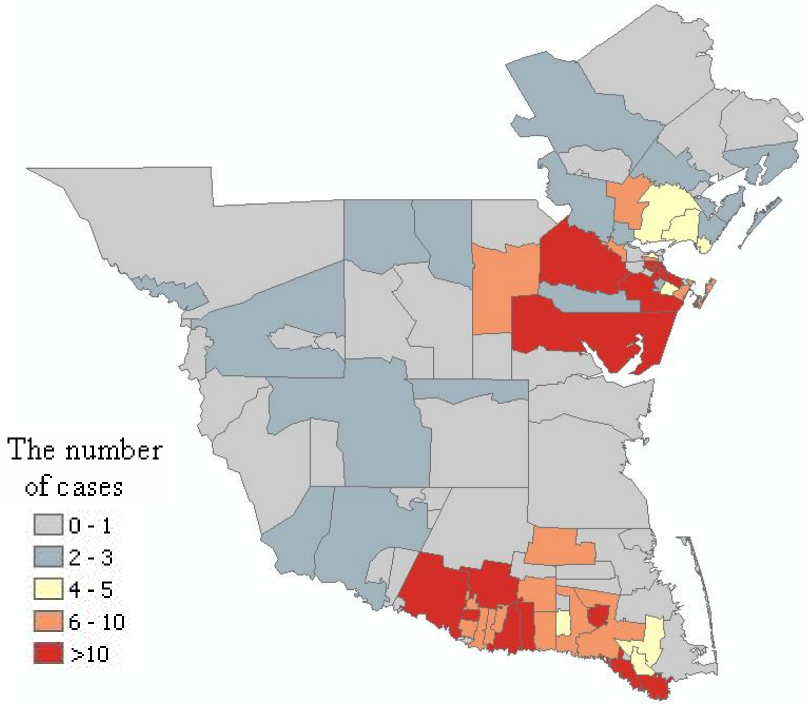


Figure 5. Distribution of murine typhus cases at the zip code level in South Texas.

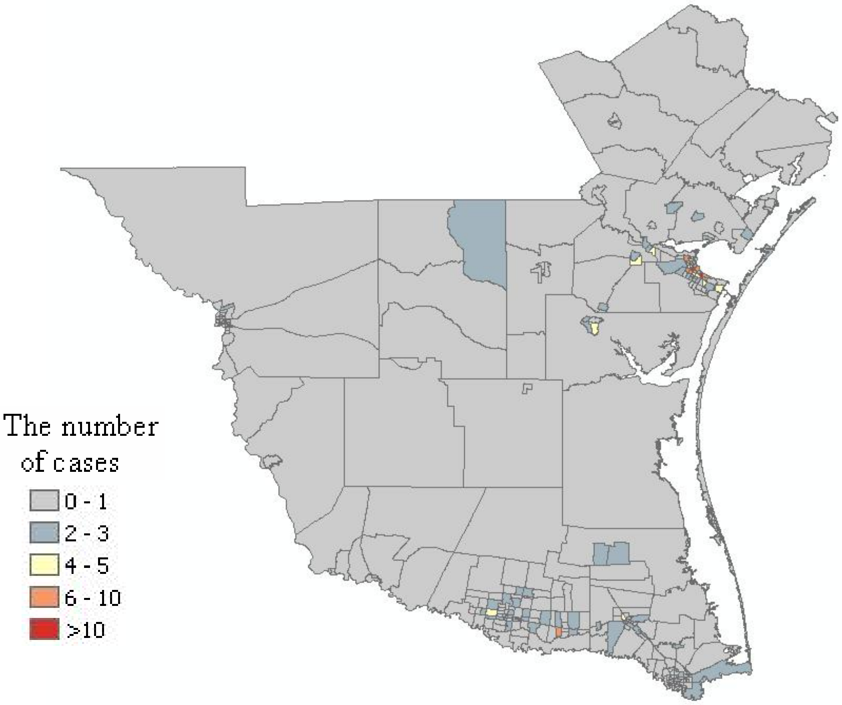


Figure 6. Distribution of murine typhus cases at the census tract level in South Texas.

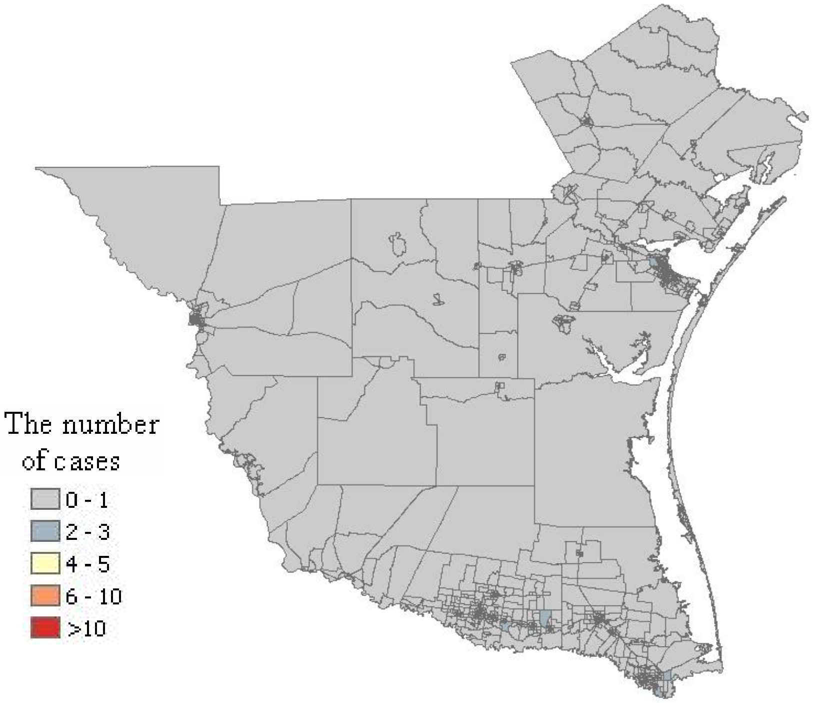


Figure 7. Distribution of murine typhus cases at the census block group level in South Texas.

4.2. Data Preparation and Processing to Support Cluster Analysis Involving Multiple Variables

Most researches about the distance proximity between the disease and environment try to explore the relationship of a single case to a single environment variable without considering the combined influences from multiple environment exposures. Our second research object was to explore the line cluster method, combined with the visualization technique, to identify the spatial correlation among multiple variables.

Regarding to the second research, the emphasis was put on the distance of each case to environmental exposures without considering the exact location of each identified disease case. The visual exploration was incorporated with the line clusters detection techniques to conduct this analysis. In order to test this method, one dataset was collected from the Texas health department in a MS excel format. The investigated group contains around 1500 identified oral clefts, organized by the case number.

Five variables were used to perform this multiple variables analysis and the values of these variables were the distance from each disease case to one variable, e.g. the near environment hazards in this research. As an example, Figure 8 shows that the investigated superfund as the potential pollution sources to the oral cleft. Since this dissertation research is a methodology research, the disease data collected in this research are just an example with no further study on the specific etiology of this disease and this proposed method could be applied to other disease studies as the method for multiple variables analysis.

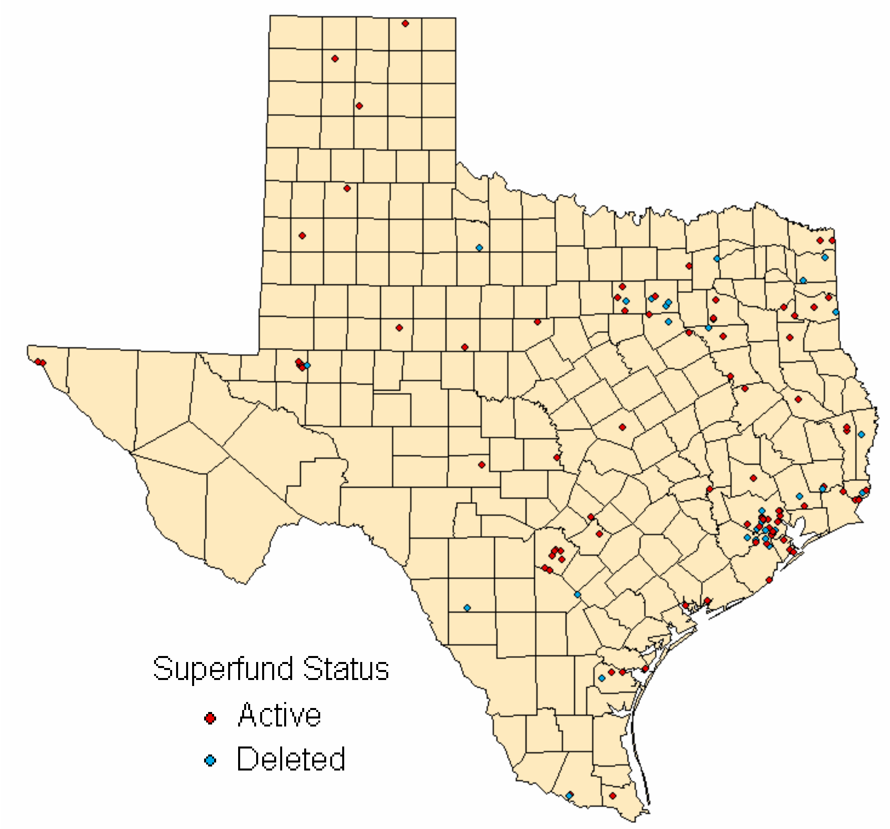
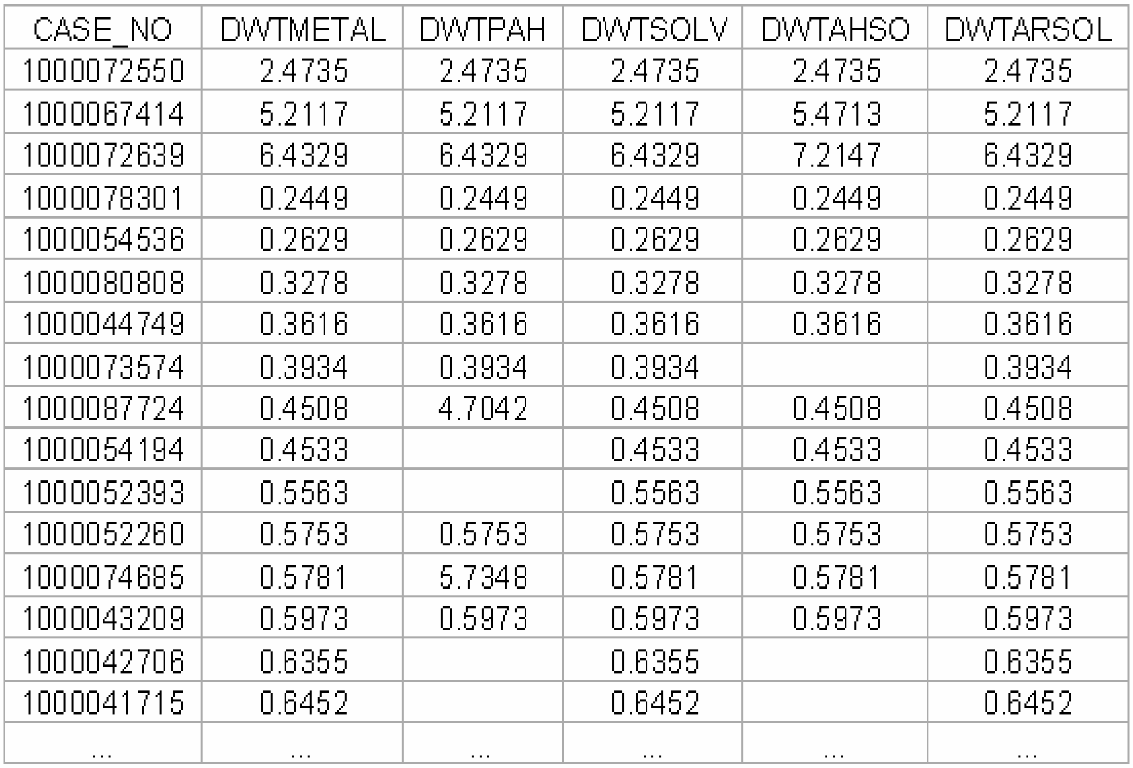


Figure 8. Superfund sites in Texas.

The data record five environmental pollution sources to the disease: the average dry weight of metal (DWTMETAL), average dry weight of polycyclic aromatic hydrocarbons (DWTPAH), average dry weight of solvent (DWTSOLV), average dry weight of aHsO (DWTAHSO), and average dry weight of aromatic solvent (DWTARSOL). The number in the table represents the distance of the identified disease case to each variable (e.g. environmental hazards) and each record represents one case. A subset of sample data is listed in the Table 1.

Table 1. The recorded diseases and distances of five closest environmental hazards



CASE_NO	DWTMETAL	DWTPAH	DWTSOLV	DWTAHSO	DWTARSOL
1000072550	2.4735	2.4735	2.4735	2.4735	2.4735
1000067414	5.2117	5.2117	5.2117	5.4713	5.2117
1000072639	6.4329	6.4329	6.4329	7.2147	6.4329
1000078301	0.2449	0.2449	0.2449	0.2449	0.2449
1000054536	0.2629	0.2629	0.2629	0.2629	0.2629
1000080808	0.3278	0.3278	0.3278	0.3278	0.3278
1000044749	0.3616	0.3616	0.3616	0.3616	0.3616
1000073574	0.3934	0.3934	0.3934		0.3934
1000087724	0.4508	4.7042	0.4508	0.4508	0.4508
1000054194	0.4533		0.4533	0.4533	0.4533
1000052393	0.5563		0.5563	0.5563	0.5563
1000052260	0.5753	0.5753	0.5753	0.5753	0.5753
1000074685	0.5781	5.7348	0.5781	0.5781	0.5781
1000043209	0.5973	0.5973	0.5973	0.5973	0.5973
1000042706	0.6355		0.6355		0.6355
1000041715	0.6452		0.6452		0.6452
...	...	...	...	...	...

CHAPTER 5

ARBITRARY SHAPE DISEASE CLUSTER DETECTION USING A NEIGHBOR-EXPANDING APPROACH

5.1. Introduction

Detecting spatial disease clusters is vital to public health surveillance. Developing tools to reveal such clusters has received considerable attention from researchers in epidemiology, mathematics, and geography, who have proposed a variety of tests to facilitate the task. As reviewed in the Chapter 2, the spatial scan statistic model proposed by Kulldorff is a widely used automatic method to detect disease cluster patterns. This method has been applied to many research fields. Examples of these applications include disease pattern analysis (Fischer et al. 2008), criminology (Minamisava et al. 2009; Nakaya and Yano 2010), network (Duczmal et al. 2007), as well as ecology and the environment (Tonini et al. 2009). However, the spatial scan statistic and other similar approaches suffer from some restrictions in practice (Neill et al. 2005; Chen et al. 2008). Although this method can be adopted to include any shape for scan windows (Kulldorff, 1997), it still has limitation in practice due to the predefined geometrical shapes of scan windows (Neill and Sabhanani 2005) which leave a large number of candidate clusters out of the test. It is therefore necessary for researchers to develop methods that can be used to detect clusters with arbitrary shapes.

Recently, many methods and strategies have been proposed to improve the detection of clusters with arbitrary shapes by constructing scanning windows of irregular shapes. Tango and Takahashi (2005) presented a “flexibly shaped spatial scan statistic” (FlexScan) which uses a limited exhaustive search to detect arbitrarily shaped clusters by aggregating its nearest circular neighboring areas (Tang and Takahashi 2005). The spatial scan statistic superimposes circular windows on the study area, while FlexScan generates irregularly shaped windows on each area by aggregating its nearest neighboring areas. To reduce the number of arbitrarily shaped scanning windows, Tango and Takahashi (2005) limited the length of clusters referring to the relatively small number of areas contained in a scanning window. This method extends the spatial scan statistic to detect irregular shapes but is only applicable for detecting clusters of small or moderate sizes. In addition, the determination of the threshold size of a cluster is very subjective, though Tango and Takahashi (2005) suggested choosing about 10~15 percent of the size of the whole study area as a reasonable number.

One solution to this problem involves setting a constraint to guide the search process so as to reduce the number of candidate scan windows. Patil and Taillie (2004) introduced the concept of “upper level set” and developed an “upper level set scan statistic”. Based on this statistic, a more generalized strategy named minimum spanning tree (also called a cheapest connecting network) was proposed by Assuncao et al (2006) to reduce the number of neighbors to be searched. This method is called a cheapest connecting network or a greedy growth search (GGS) which only absorbs the neighboring areas to maximize the likelihood of a new window. This idea was further improved in the Density-Equalizing Euclidean Minimum Spanning Tree (DEEMST) method proposed by Wieland and her colleagues (2007). The Minimum Spanning Tree method offers two different functions: in a static minimum spanning tree, the weight refers to the difference of risk rate; in a dynamic minimum spanning tree, the variance of maximum likelihood ratio is taken into account. These methods are similar to GGS as they absorb only the neighboring areas in the search process to maximize the likelihood of a new window. It has the flexibility to start the search from any location in the study area.

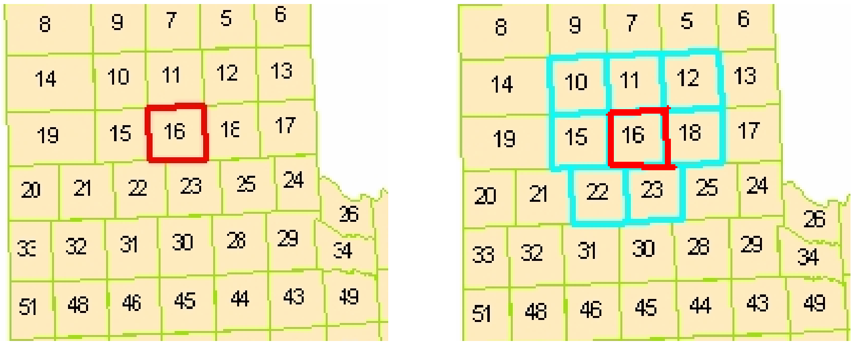
GGS cannot avoid the local maximum problem (Duczmal and Assuncao 2004). Many algorithms were adopted or developed to improve the GGS. The genetic algorithm is employed to limit the irregular shape of most potential real clusters (Conley et al. 2005; Sahajpal et al. 2005; Duczmal et al. 2007). Yiannkoulias et al. (2007) presented two approaches to improve the greedy growth search: one is the non-connectivity penalty in order to limit the very irregular cluster shapes and the other is the depth limit (*u*) to prevent the generation of large super-clusters from smaller clusters (Yiannakoulias 2007). These approaches will terminate the search in GGS if it fails to increase the likelihood after some steps.

Another famous improvement is a “simulated annealing strategy” proposed by Duczmal and Assuncao (2004). This method is based on graph theory in which nodes present centers of areas, and edges present the geographical relationships among areas (Duczmal and Assuncao 2004). The simulated annealing spatial scan statistic was improved by introducing a non-compactness penalty to reduce the chance that the cluster with extremely irregular shapes would be found (Duczmal et al. 2006). Most of the recent proposed methods try to detect the globally most likely cluster (Duczmal and Assuncao 2004; Duczmal et al. 2007) and this is critical in cluster detection since the search process of some methods frequently leads to or sticks on the locally most likely clusters.

In this chapter, the development of two algorithms were reported that use a new neighbor-expanding approach based on the assumption that any subset of adjacent areas could make up a potential cluster, and that the shape of this cluster might not be circular or rectangular. These two algorithms are called the maxima -likelihood-first (MLF) algorithm and non-greedy growth (NGG) algorithm. These two algorithms build upon the existing cluster detect techniques, and adopt neighbor-expanding tactics to construct a set of scan windows instead of just using the scan windows in some predefined shapes. Furthermore, the proposed algorithms improve the arbitrarily-shape cluster detection method in avoiding the local maximum problem since the algorithms search for the globally most likely cluster at each step in the search process.

5.2. A New Neighbor-expanding Approach

A new neighbor-expanding approach was developed here to detect clusters with arbitrary shapes. Suppose we have a map consisting of a tessellation of component areas. These areas are associated with case numbers and the total population at risk. Two areas were considered as neighbors when their boundaries are touched. It was assumed that a region with any set of connected areas may make up a potential cluster and a cluster may appear in different shapes depending on how many and how aggregated the set of connected areas are. The goal is to find such clusters with the likelihood ratio in the scan statistic. In the search process, a large subset of connected areas was swept, constructing a new region at each step by aggregating one of its neighbor areas, until certain thresholds were met or the expected results were obtained. For the sake of simplicity, the length was used to indicate the number of areas that constitute a region. Usually, a new region is derived with a higher length ***k+1*** by combining a ***k*** length region and one of its neighboring areas. One can easily figure out the number of regions with ***k+1*** length based on a ***k*** length region. If the number of the neighbors around ***k*** length region is ***j***, then one can obtain ***j*** regions at ***k+1*** length. To clarify, this process is illustrated using an example as shown in Figure 9. In Figure 9, every area is labeled with a number on it. A set of numbers were used to represent the region that is made up of both a region and its neighbors. For example, {16} means a region containing a single area 16 and {16, 18} corresponds to a region consisting of areas 16 and 18.

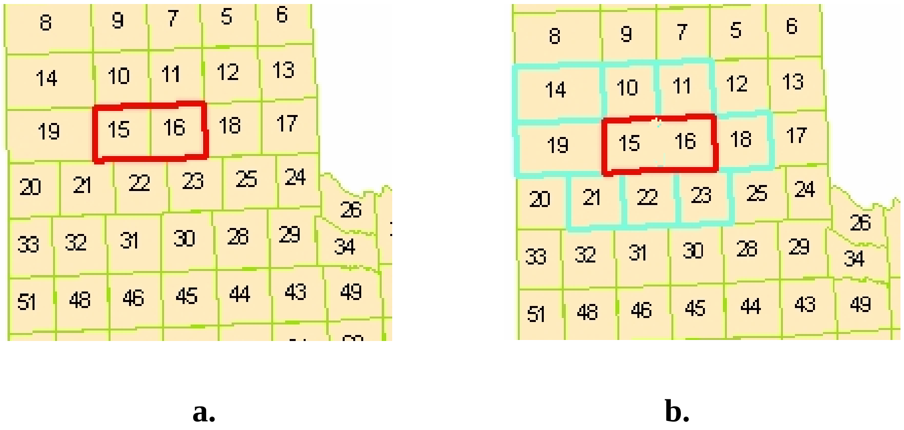


a. **b.**

Figure 9. Example of neighbor-expanding. a) an example map showing a chosen region; b) the neighbor areas. The red color highlights the chosen area and cyan color highlights the neighbor areas.

Now if {16} is a seed region at first length, which is highlighted by red color in Figure 9a, and it has seven neighbors, area 10, 11, 12, 15, 18, 22, and 23. The neighbors of {16} are shown in the Figure 9b. Thus the seven regions can be obtained at the second length based on region {16}. These seven regions are {10, 16}, {11, 16}, {12, 16}, {15, 16}, {18, 16}, {22, 16}, and {23, 16}. Furthermore, in order to obtain the third length regions, the region {15, 16} has the following neighbor areas: 14, 10, 11, 12, 13, 19, 18, 17, 21, 22, and 23. Region {15, 16} and its neighbor areas are illustrated in Figure 10b.

Now 11 regions are derived at the third length: {14, 15, 16}, {10, 15, 16}, {11, 15, 16}, {12, 15, 16}, {13, 15, 16}, {19, 15, 16}, {18, 15, 16}, {17, 15, 16}, {21, 15, 16}, {22, 15, 16}, and {23, 15, 16}.



a. **b.**

Figure 10. Example of neighbor-expanding. a) region {15, 16}; b) the neighbor areas of region {15, 16}. Using red color to highlight the chosen area and cyan color to highlight the neighbor areas of the chosen area.

While this search process continues, the number of regions increases exponentially as we aggregate more areas. This process is computationally very intensive. In order to reduce the number of regions, two alternative algorithms were developed for the construction of regions or scan windows: maxima-likelihood-first (MLF) algorithm and non-greedy growth (NGG) algorithm.

5.2.1. Maximum-likelihood-first Algorithm

The principal goal of this algorithm was to direct the new region construction process to obtain a global maximum. This maximum refers to the highest value obtained by the proposed approach. After analyzing equations (4) and (5) in Chapter 3, it was found that it was hard to determine which of the following factors make the most contribution to the likelihood ratio: the number of cases, population size, or the relationship between them. Thus, there is no clear guidance that could help to construct scan windows which would have the highest likelihood ratios. Rather than construct scan windows randomly, the focus of this algorithm is to generate windows for the most promising clusters. This approach was named as the maximum-likelihood-first (MLF) approach because it always constructs new promising clusters by expanding from the current best candidate, yielding the maximum likelihood ratio.

The proposed approach is illustrated in the flowchart in Figure 11. In the initial step of the algorithm, the Log likelihood ratios (LLRs) are calculated for all areas and put the elevated LLRs into a temporary candidate list. After sorting their LLRs in the temporary candidate list, the one with the highest LLR is selected as the candidate region. In the next step, the candidate region aggregates one of its neighboring areas to create a new region. A group of new regions are obtained and the LLRs of these new regions are calculated. These new regions are put into the temporary candidate list, and then the new and old members are sorted in the candidate list together again, and the one with the new maximum LLR is selected as the new candidate. Unlike the minimum spanning tree algorithm (Assuncao et al. 2006), this algorithm expands the neighbors based on multiple seeds in the cluster candidate list. The seed for each neighbor expansion is selected from all the candidates in the temporary candidate list. The procedure is repeated until either the aggregated area covers half of the study area or has half of the total population.

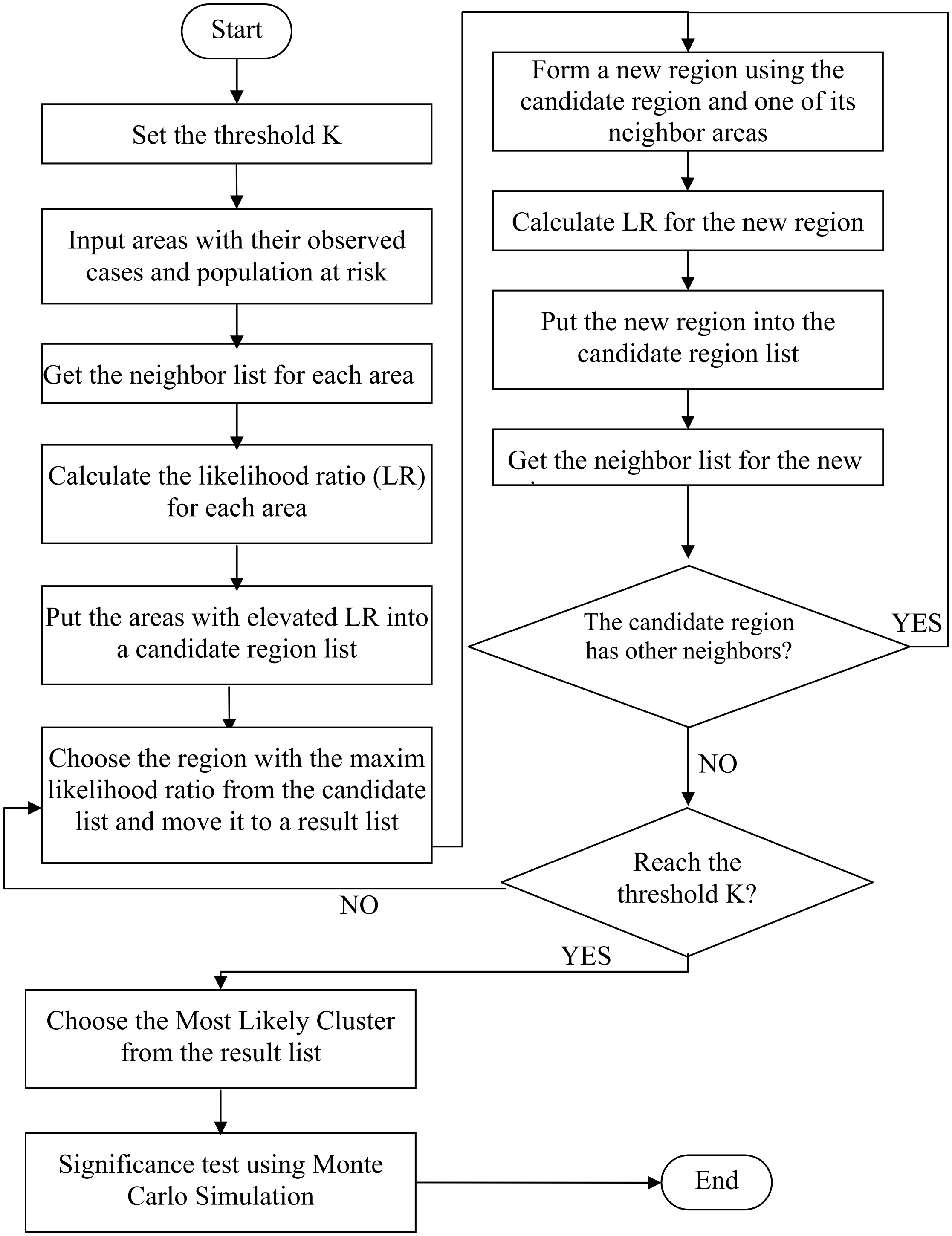

Start
Set the threshold K
Input areas with their observed cases and population at risk
Get the neighbor list for each area
Calculate the likelihood ratio (LR) for each area
Put the areas with elevated LR into a candidate region list
Choose the region with the maxim likelihood ratio from the candidate list and move it to a result list
Form a new region using the candidate region and one of its neighbor areas
Calculate LR for the new region
Put the new region into the candidate region list
Get the neighbor list for the new
The candidate region has other neighbors?
YES
NO
Reach the threshold K?
NO
YES
Choose the Most Likely Cluster from the result list
Significance test using Monte Carlo Simulation
End

Figure 11. A flowchart illustrating the maximum-likelihood-first algorithm.

When detecting the cluster using the neighbor-expanding approach described above, it is very likely that the procedure may stick to some areas with high LLRs and unable to search the entire study area. Usually, LLRs of candidate clusters depend on the risk rates of their neighbors (Wieland et al. 207). That is, areas with higher risk rates are more likely to have higher LLRs than those with lower risk rates since LLRs of clusters do not vary a lot if they contain the same subset of areas (Kulldorff 1997). It means if a candidate cluster overlaps largely with another candidate cluster with a high LLR, it may have a higher LLR than other areas which have not been explored. This observation leads to proposed search procedure to stick with one area and its neighbors if their LLRs increase fast at the beginning and decrease slowly. Therefore, it is necessary to set a threshold to stop the search around a particular area and its neighbors when the LLRs of the newly generated clusters fail to increase in certain steps. This arrangement allows the search to move to other unexplored areas to detect other potential cluster centers. Originally suggested by Yiannakoulias, Rosychuk, and Hodgson (2007) as a depth limit adaptation, this idea is incorporated into the MLF algorithm.

As shown in Figure 11, this procedure was repeated until half of the total population or study area is covered. The cluster with the highest LLR was selected as the most likely cluster while the secondary cluster is the cluster having both the second highest LLR with no overlap area with the most likely cluster. Since this approach does not focus on one or some particular areas, it is expected to avoid the local maximum problem.

5.2.2. Non-greedy Growth Algorithm

The non-greedy growth (NGG) algorithm is an improved version of greedy growth algorithm (Yiannakoulias 2007). Several researchers have described how greedy growth approaches perform in searching clusters with irregular shapes (Duczmal et al. 2006; Yiannakoulias 2007). The greedy growth search starts with areas having high log likelihood ratio as seed areas for potential clusters. The search is only interested in a neighboring area that has the maximum LLR or has the capability to maximize the LLR when aggregated to form a new potential cluster. Similar to the procedure described above, the greedy growth algorithm joins other areas until a given population size or other thresholds are reached. The same procedure is repeated from other seed areas.

The greedy growth approach sounds tempting, but it has an inherent deficiency in that it does not guarantee to find either the best solution or the global maximum. This method easily falls into the trap of local maximum since it excludes some areas which might potentially form a more promising cluster when they combine with other areas.

To solve this problem, a new algorithm was proposed to minimize the impact of the local maximum problem. To distinguish it from traditional greedy growth approaches, we name it “the non-greedy growth algorithm”. The algorithm allows not only the neighboring area with the local maximum to be included but also includes many other neighboring areas in the search procedure. Usually the number of newly formed regions relies on the number of candidate regions and the number of neighbors of each region. With this method, a constraint can been set on each of these two numbers control the number of newly formed regions at the next step of the search process. Previous studies suggest that the number of candidate regions increase exponentially, while the number of neighbors of each region does not change dramatically. Therefore, it is more reasonable to set a threshold on the number of candidate regions. Theoretically, if only one candidate and one of its neighbors were chosen with the highest LLR each time, this method degrades to the traditional greedy growth search method. The inverse extreme of this approach is the naïve exhaustive approach where no limitation is set.

In the NGG algorithm, a threshold (M) is set on the maximum expected number of new regions in each iteration. Given that threshold and the average number of neighbors, it is easy to determine how many candidate regions should be chosen to participate in the aggregation process. There are a few options in the choice of candidate regions. One is to choose M most promising regions, directly from the pool of candidates, or to choose them randomly. In the actual implementation reported in this chapter, a combination of the two was used, that is, part of M candidates are from the top regions and the rest are chosen randomly.

The flowchart showing the NGG algorithm is given in Figure 12. At first, a threshold M is set for the maximum number of potential clusters generated at each step. Then all areas are put into a temporary list and the LLRs of these areas are calculated. In the next step, the average number of neighbors (L) of each region is calculated. The approximate number of candidates (N) for the next iteration is estimated by the preset parameter M and the average number of neighbor L using the equation N = M/L. N areas with the highest LLRs were chosen from the temporary list and the list was emptied afterward. New regions created from the candidates and their neighbors were put into the

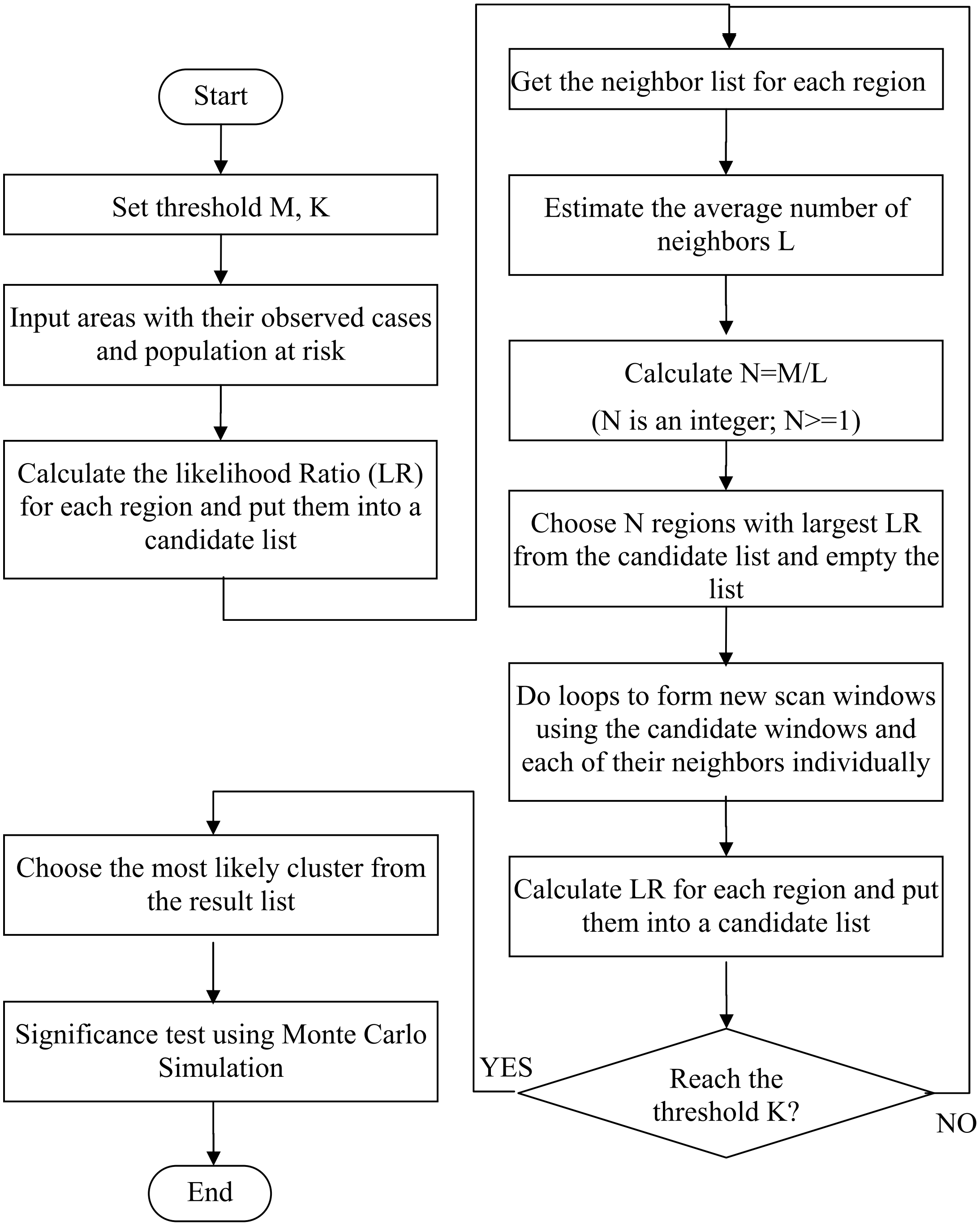

Start
Set threshold M, K
Input areas with their observed cases and population at risk
Calculate the likelihood Ratio (LR) for each region and put them into a candidate list
Get the neighbor list for each region
Estimate the average number of neighbors L
Calculate N=M/L
(N is an integer; N>=1)
Choose N regions with largest LR from the candidate list and empty the list
Do loops to form new scan windows using the candidate windows and each of their neighbors individually
Calculate LR for each region and put them into a candidate list
Reach the threshold K?
YES
NO
Choose the most likely cluster from the result list
Significance test using Monte Carlo Simulation
End

Figure 12. A flowchart showing the non-greedy algorithm.

emptied list. These steps were repeated until either the aggregated area covers half of the study area or has half of the total population.

MLF and NGG have their own advantages and disadvantages. An comparison of these two algorithms is presented below.

Table 2. An initial comparison of MLF and NGG

	Advantage	Disadvantage	Favored Situation
MLF	• results might be more significant with higher LLRs • it is faster than NGG when there are few clusters	• it is hard to control when most clusters have relative similar LLRs • only the cluster with the highest LLR is kept into the next search	• data containing few extreme clusters • small number of units
NGG	• the maximum number of candidate cluster is controllable • it is simple to be implemented	• the search procedure will continue until it reaches the criteria	• large number of units

5.3. Results and Discussion

5.3.1. Performance Test Using Simulated Data and Benchmark Data

The performance of the two new algorithms was evaluated and compared with the simulated annealing (SA) strategy method, flexible-shape scan statistic (FlexScan), and spatial scan statistic (SaTScan) before applied to the south Texas data. The simulated data consisted of a tessellation of approximately 300 hexagon component areas (Figure 13). These hexagonal areas had the same size. It was assumed that populations were homogeneously distributed, and that each hexagonal area had an equal population (1000 persons) subject to disease risk. The areas falling in a synthesized cluster were assumed to have a high risk rate of 0.5% (5 cases /1000 person) while areas outside have a low risk rate of 0.2% (2 cases / 1000 person). The comparisons were based on five different scenarios: a compacted cluster, a ring-shape cluster with regular patterns, an elongated-shape cluster, a strange-shape cluster, and a two-shape cluster with irregular patterns.

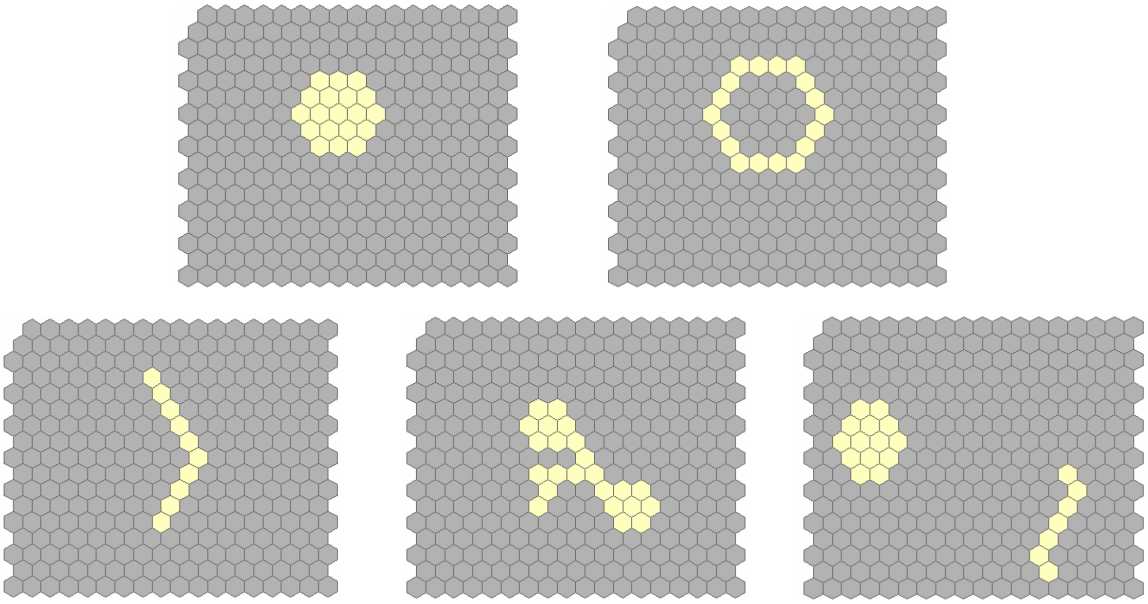


Figure 13. The simulated five cluster patterns for the performance test.

Figure 14 shows the most likely and secondary clusters detected by the MLF, NGG, SA, FlexScan, and SaTScan. Our methods, both MLF and NGG, and SA performed better than the FlexScan and SaTScan methods. Obviously, the SaTScan only performed very well on the compact regular cluster, achieving the same LLR and p-value as other methods (Table 3). However, as the pattern became less regular or less compact, the performance of SaTScan became unsatisfied. The worst performance was found in the two-cluster pattern, with the largest p-value (0.998) and the smallest LLR value (2.627). The FlexScan method did not perform well in situations involving the ring shape or two-cluster shape with small LLRs (7.165 and 6.599) and large p-values (0.836 and

0.954). The possible reason is that the FlexScan method tries to search for the nearest neighbor; this strategy would trap the search at a location since most of neighbors in the ring and two-cluster patterns are far away from each other. For the extreme irregular shaped patterns, two sub-clusters were detected by the SaTScan with a much less LLR value (9.143) than that of the MLF (32.513). With the two-cluster

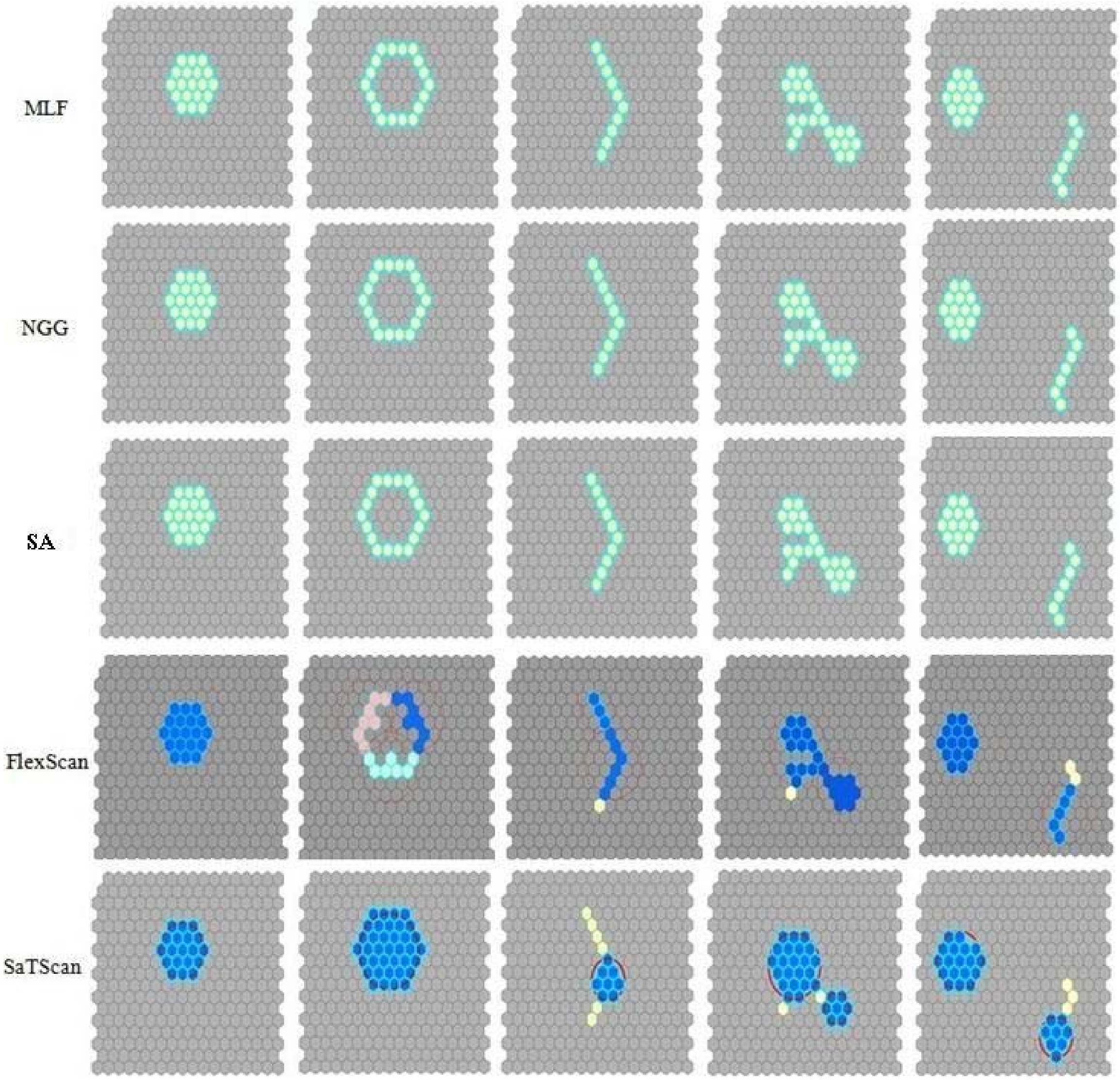


Figure 14. The first line shows the most likely and secondary clusters detected by the spatial scan statistic and the second line shows the most likely and secondary clusters detected by the maximum-likelihood-first method.

pattern, the secondary cluster shows much weaker in the SaTScan method with a larger p-value (0.998) and a smaller LLR (2.627). These results indicate that SaTScan and FlexScan are not appropriate in catching clusters with irregular shapes.

Table 3. The comparison between the MLF method, NNG method, SA method, Tango's FlexScan method and Kulldorff's SaTScan method using the synthesized data

Clusters		Observed #	Expected #	LLR	p-value
Compact shape	MLF	95	41.646	27.396	0.001
	NNG	95	41.646	27.396	0.001
	SA	95	41.464	27.396	0.001
	FlexScan	95	41.646	27.396	0.001
	SaTScan	95	41.646	27.396	0.001
Ring shape	MLF	90	39.273	26.083	0.001
	NNG	90	39.273	26.083	0.001
	SA	90	39.273	26.083	0.001
	FlexScan	32	15.273	7.165	0.836
	SaTScan	128	80.730	13.756	0.001
Long shape	MLF	50	21.010	15.069	0.001
	NNG	50	21.010	15.069	0.001
	SA	50	21.010	15.069	0.001
	FlexScan	30	12.606	8.866	0.432
	SaTScan	28	16.810	3.202	0.993
Extreme shape	MLF	115	51.343	32.513	0.001
	NNG	115	51.343	32.513	0.001
	SA	115	51.343	32.513	0.001
	FlexScan	65	29.020	17.477	0.003
		45	20.091	11.877	0.081
	SaTScan	86	49.110	12.425	0.001
		35	15.630	9.143	0.024
Two-cluster	MLF	70	30.970	19.367	0.001
		35	15.485	9.343	0.016
	NNG	70	30.970	19.367	0.001
		35	15.485	9.343	0.016
	SA	70	30.970	19.367	0.001
		35	15.485	9.343	0.016
	FlexScan	70	30.970	19.367	0.001
		25	11.061	6.599	0.954
	SaTScan	78	39.820	15.470	0.001
		28	17.700	2.627	0.998

A further comparison was performed among these methods using the benchmark real disease data. The data were collected from 11 states and the District of Columbia in the Northeast US from 1988 – 1992, consisting of 58,943 deaths from breast cancer among women. Figure 15 shows the most likely clusters detected by MLF, NGG, SA, FlexScan, and SaTScan methods and Table 4 summarizes these results. For the most detection methods, the most likely clusters had significantly lower p-values (≤ 0.01) and high LLR values (Table 4). Based on the p-value and LLR values, it was concluded that MLF is the most accurate method for detecting clusters with arbitrary shapes, followed in decreasing order by SA, NGG, FlexScan, Elliptic SaTScan, and Circular SaTScan.

Table 4. A comparison of the MLF method, NNG method, Duczmal's SA method, Tango's FlexScan method, and Kulldorff's SaTScan method using the benchmark data

	MLF	NGG	SA	FlexScan	SaTScan	
					Circular	Elliptic
Population	29,535,210					
Total case	58,943					
Observed #	17,002	17,743	15,122	6,980	21,039	15,122
Expected #	14,166	15,383	12,988	6,005	19,734	12,988
LLR	237.24	85.97	227.11	84.11	44.95	44.71
p-value	0.001	0.001	0.001	0.001	0.01	0.001

Note: # means number; LLR means log-likelihood ratio.

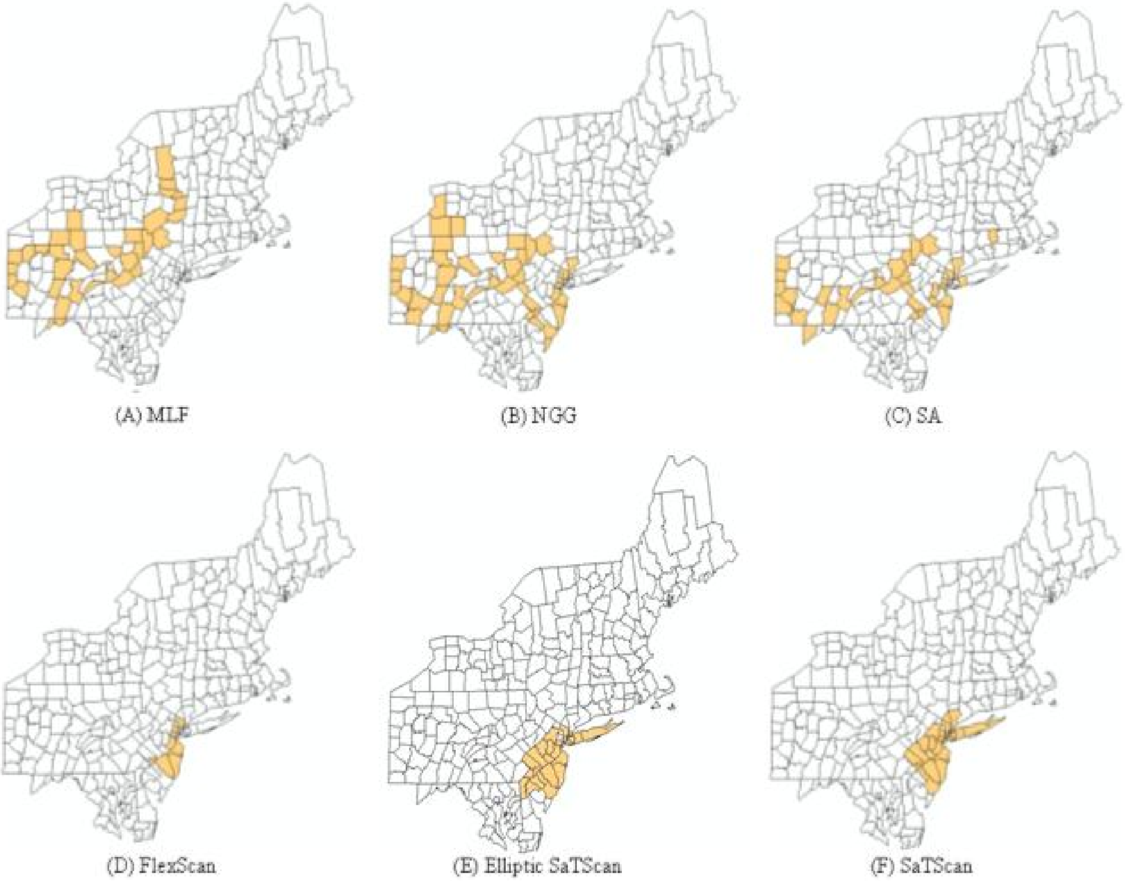


Figure 15. The most likely cluster in the benchmark real disease data detected by MLF, NGG, SA, FlexScan, and SaTScan.

5.3.2. Detection of Cluster with Arbitrary Shapes

The spatial distribution of murine typhus in the south Texas from 1998 – 2008 was identified using the new neighbor-expanding approach developed in this study and traditional SaTScan, FlexScan, and SA methods. The most likely clusters and the secondary clusters detected by the methods are showed in figure 16 (MLF), Figure 17 (NGG), Figure 18 (SA), Figure 19 (FlexScan), Figure 20 (Elliptic SaTScan), and Figure 21 (Circular SaTScan). Both the most likely clusters and the secondary clusters detected by these six methods are highlighted.

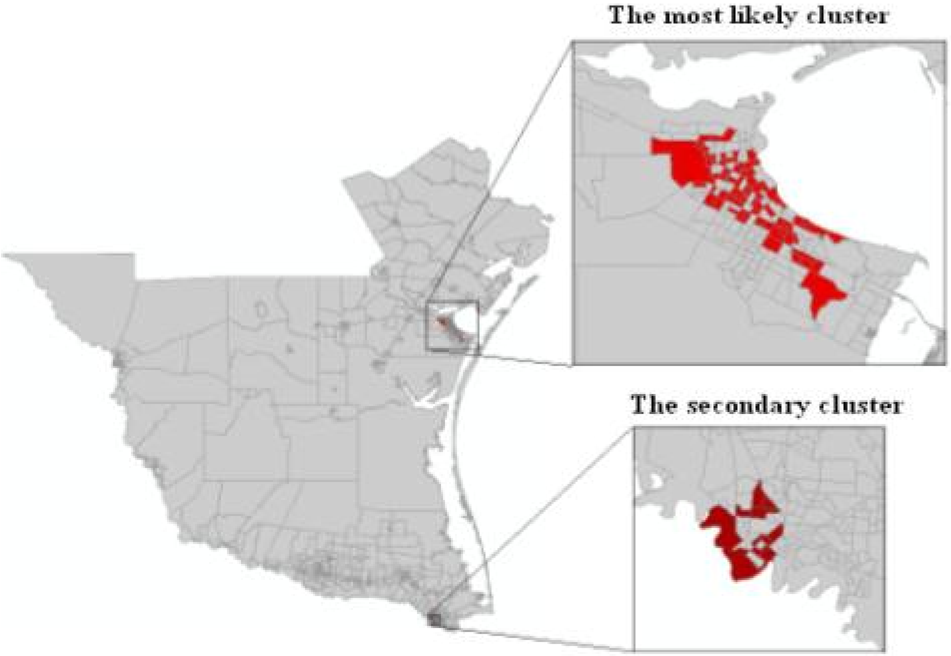


Figure 16. The most likely cluster and the secondary cluster detected by the MLF method at the census block group level.

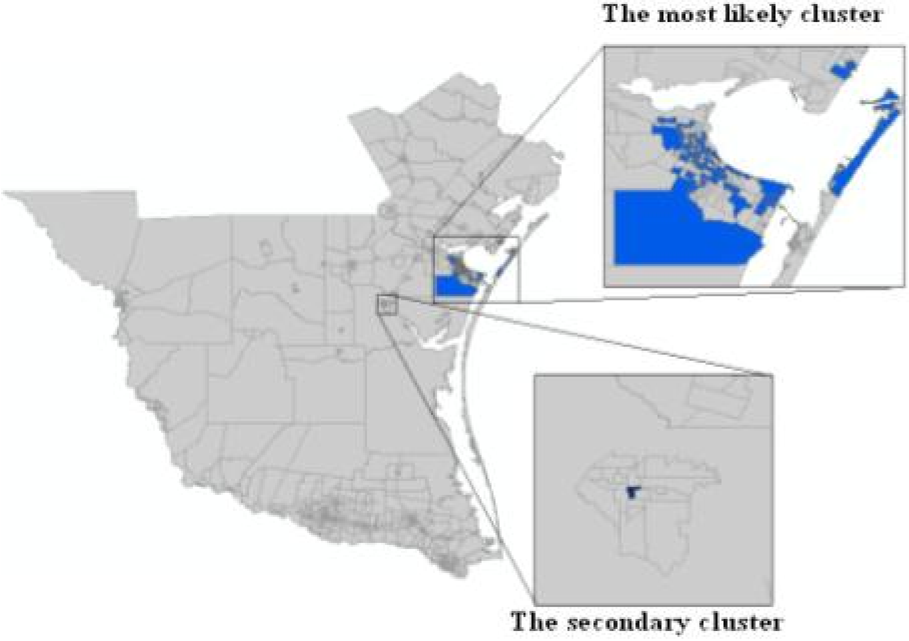


Figure 17. The most likely cluster and the secondary cluster detected by the NGG method at the census block group level.

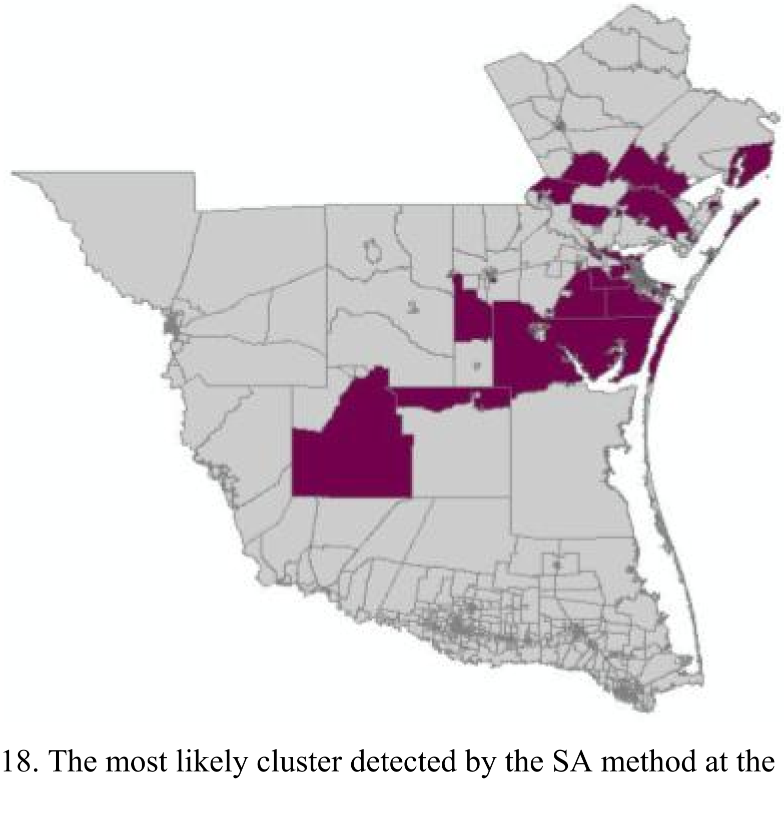


Figure 18. The most likely cluster detected by the SA method at the census block group level.

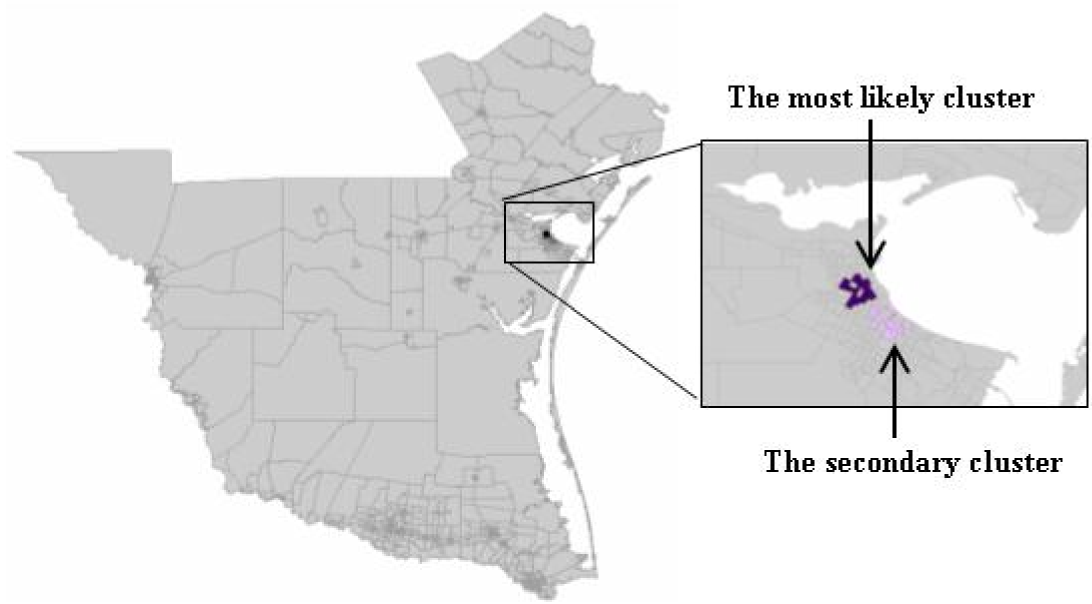


Figure 19. The most likely cluster and the secondary cluster detected by the FlexScan method at the census block group level.

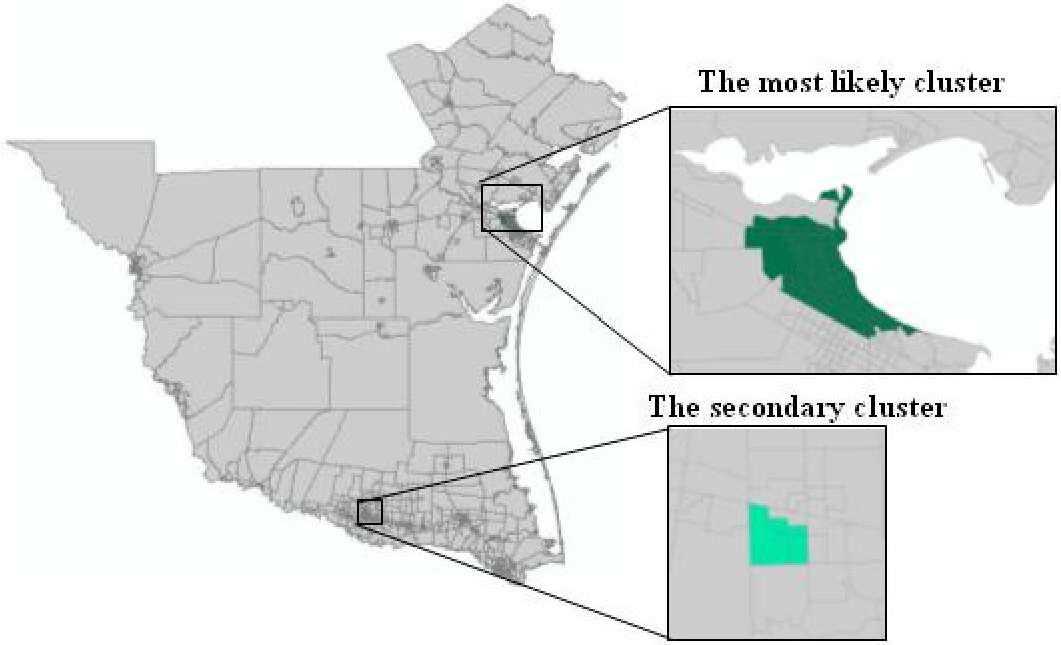


Figure 20. The most likely cluster and the secondary cluster detected by the Elliptic SaTScan method at the census block group level.

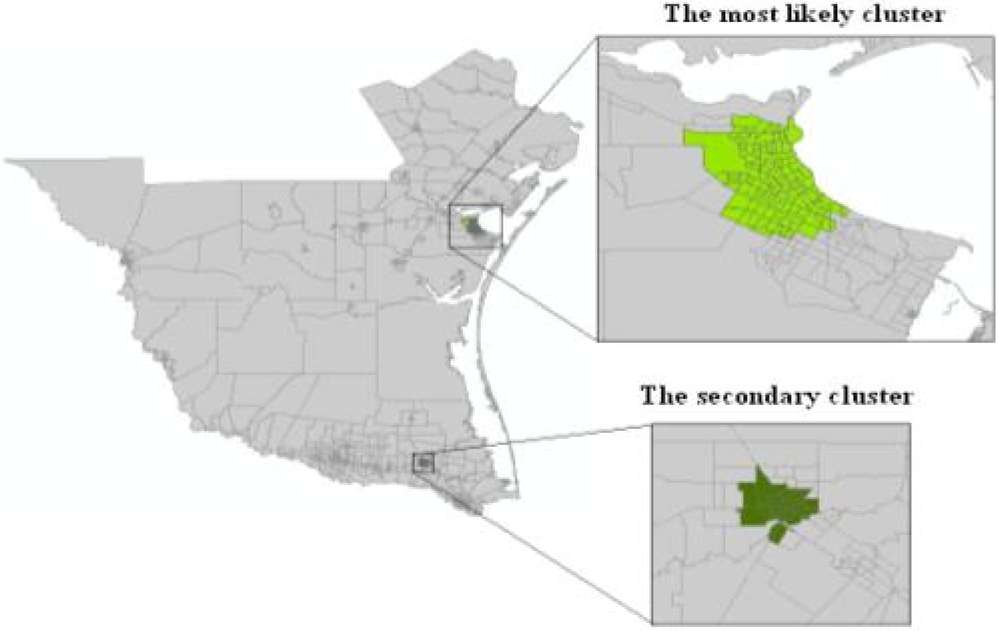


Figure 21. The most likely cluster and the secondary cluster detected by the Circular SaTSCan method at the census block group level.

Table 5. Results of cluster analysis of the Murine Typhus case in south Texas from 1996 to 2000 at the census block group level

	MLF		NGG		FlexScan		SaTScan				SA	
	Most Likely Cluster	Secondary Cluster	Most Likely Cluster	Secondary Cluster	Most Likely Cluster	Secondary Cluster	Most Likely Cluster		Secondary Cluster		Most Likely Cluster	Secondary Cluster
							Circular	Elliptic	Circular	Elliptic		
Population	1,728,393											
Total case	391											
LLR	186.43	9.33	197.51	6.15	42.95	36.95	97.60	124.69	6.67	6.49	177.15	N/A
# of zones	71	11	94	1	16	9	127	121	27	3	164	N/A
Observed #	142	12	167	3	30	25	145	138	2518	6	220	N/A
Expected #	18.96	2.53	26.5	0.15	3.01	2.37	33.54	28.99	6.69	0.87	50.5	N/A
p-value	0.01	0.25	0.01	0.32	0.01	0.01	0.01	0.01	0.42	0.74	0.01	N/A

As shown in the figures, all the most likely clusters found by the algorithms are significant with a p-vaule of 0.01 and high LLR values. The LLR value of the most likely cluster detected by the MLF algorithm (186.43) and NGG algorithm (197.51) are slightly higher than that of the SA algorithm (177.15) and significantly higher than that of the FlexScan method (42.95) and Circular SaTScan (97.60) (Table 5). The number of most-likely clusters detected by the NGG method (94) is obviously larger than that from the MLF method (71) while the number of secondary clusters detected by the NGG method (1) is much less than that from the MLF method (11). A possible reason for this result is the design of the algorithm itself. Instead of finding the maximum value in the candidate cluster, the NGG algorithm keeps expanding to its neighbors by selecting multiple candidates as seeds for subsequent steps. This procedure will surely lead to a wide distribution of the most likely clusters. Another significant difference found in the NGG algorithm is the shape of detected clusters. Although the distribution of detected clusters is very similar, it was found that the shape of clusters detected by the NGG algorithm (Figure 17) is more irregular than that from the other three algorithms. The potential reason is the same: the algorithm keeps expanding to its neighbors by selecting multiple candidates as seeds for next steps. Since we did not incorporate any penalty function to restrict neighbor expanding, it will influence the direction of the search and the power of the NGG algorithm significantly (Duczmal et al. 2006).

5.3.3. Spatial Distribution of Cluster and Socioeconomic Factors

An examination of Figures 16-21 reveals that the presence of the most likely clusters is mainly distributed in the coastal counties, particularly in Nueces County. Caused by two organisms, *Rickettsia typhi* and *R. felis* [Azad 1990], murine typhus is easily carried and transmitted by small mammals such as mice, domestic cats, and opossums and the associated fleas. Theoretically, the spreading of murine typhus requires a warm and humid environment. This is probably why most of the detected clusters are distributed in the coastal area.

The distribution of population and related environmental problems might be the reasons responsible for clustering of the cases. Figure 22 is the population density at the census block group level. Of the total 1,068 census blocks in the study area, half of them (534) have more than 1,000 persons per square kilometer. Most of these density populated counties are found in the eastern coastal region and in the southern area. The large cities in the southern area are the city of McAllen and Brownsville, and the largest city in the eastern coastal region is Corpus Christi. Not surprisingly, these large cities with high population densities are the major seating area of the detected most likely cluster and secondary clusters in this study. In the MLF method, there are 71 census block groups detected out as the most likely cluster and 66 of them (92.96%) had densities higher than 1,000 persons per square kilometers; 42 of them (59.15%) had densities higher than 2,000 persons per square kilometers (Table 6). A similarly high percentage could be found in FlexScan (100%), Circular SaTScan (91.34%), and Elliptic SaTScan (90.18%).

The similarity between the distribution of cluster patterns and the environmental factors can be found for all the cluster detection results. Most of reported cases are found in urban areas with very high population densities. Usually, the high density population brings problems, such as increasing amounts of urban garbage and commensal rodents. These will also increase the likely exposure of opossums, a peridomestic animal, to the cat fleas and rickettsial pathogens due to their frequent visiting of human habitation to search for both food and harborage (Wen and Kedem 2009). Moreover, the high population densities also enlarge the number of household pet, which is another common host of cat fleas. Besides the rats and mice, the cat flea is easily switched from the parasitized cats and opossums to other animals of the same size.

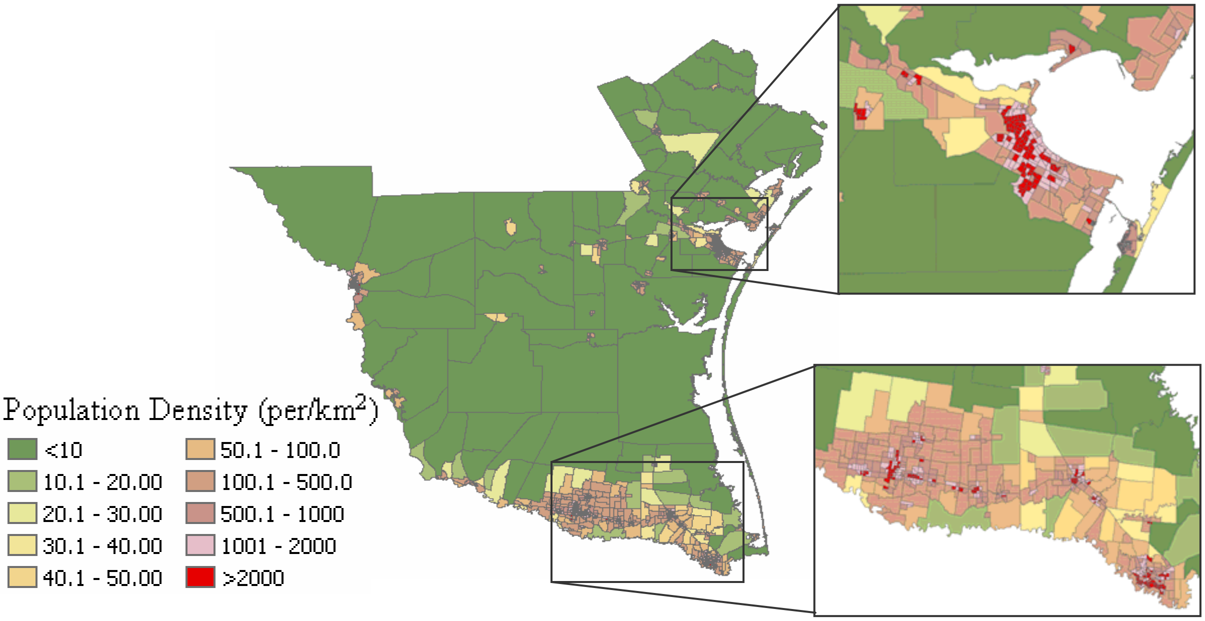


Figure 22. The population density at the census block group level.

Table 6. Relation between the number of most likely cluster and areas with high population density

	# of cluster	Density > 1000		Density > 2000	
		# of cluster	Percentage (%)	# of cluster	Percentage (%)
Max-First	71	66	92.96	42	59.15
Non Greedy	94	77	81.91	44	46.81
SaTScan	127	116	91.34	67	52.76

To further verify and explain the detected cluster patterns, four other socioeconomic factors were collected and analyzed at both county level and census block group level: median household income, the rate of population with their poverty status below poverty, median house built year, and median value of owner-occupied house units. Nueces County, with the majority of the most likely clusters, has a relative higher median household income ($35,959) and median house value ($70,100) than the average value (median household income $27,026 and median house value $48,467) for all 18 counties. Driven mainly by tourism and the petrochemical industry, the main economic support of Nueces County depends upon its largest coastal city, Corpus Christi, which also drives the development of related commercial real estate and other industries.

For the socioeconomic analysis at the census block group level, the location and distribution of the most likely clusters detected by MLF, NGG, SA, FlexScan, Elliptic SaTScan, and Circular SaTScan within Nueces County (Figure 23) and the associated socioeconomic data (Table 7) were illustrated. Compared to the average value of all block groups within Nueces County, the median household income and house value of the ‘clustered’ census block groups are obviously lower than those in other block groups. Meanwhile, the poverty rate of this ‘hot spot’ area is relatively higher than the average poverty rate in all of Nueces County. All these data indicate that the detected cluster patterns agree with the socioeconomic distribution which plays a critical role in the transmission of murine typhus. It is also likely that other information, such as the habitual environment of human and city animals, as well as transmission among people, may be critical in tracking the transmission model. This would be another interesting topic of future research if ancillary data can be obtained in the future.

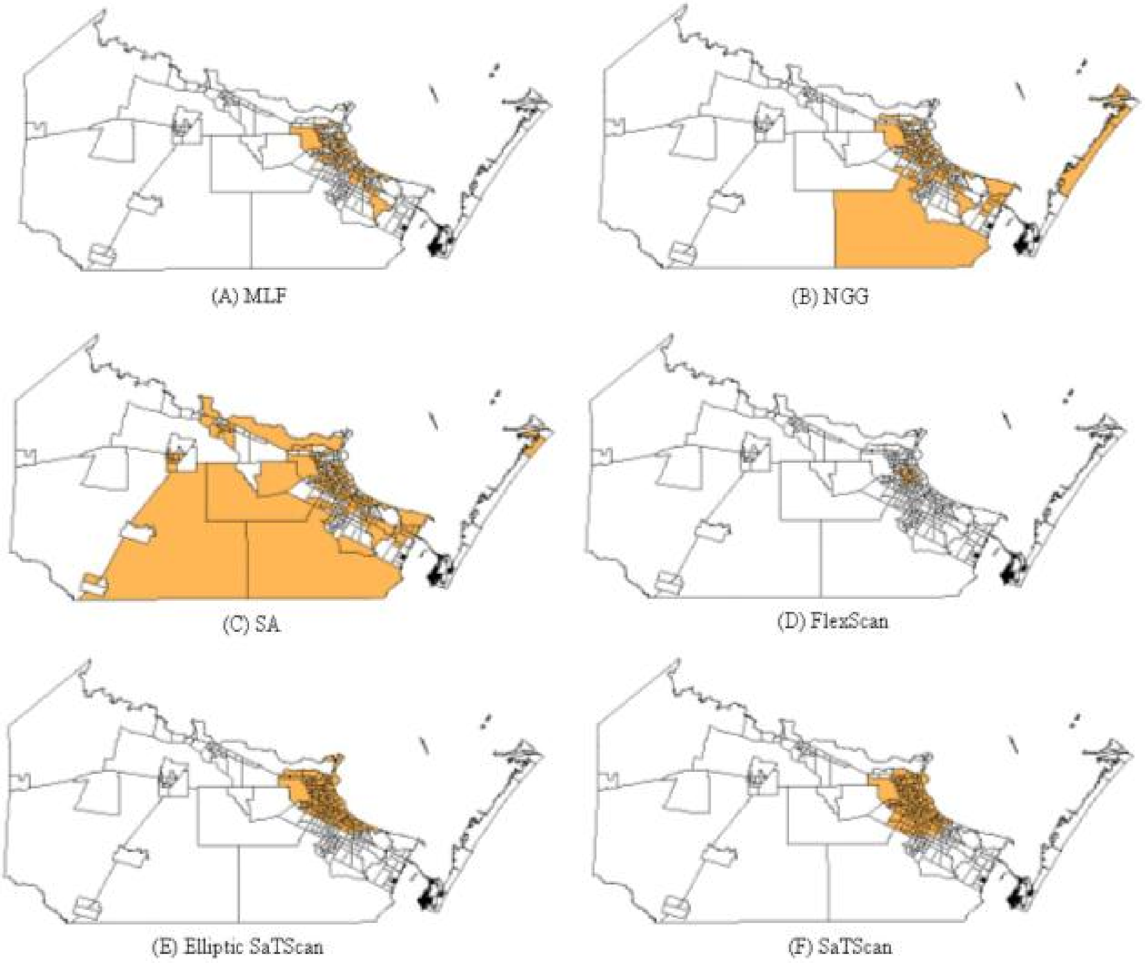


Figure 23. The most likely cluster detected within the Nueces County.

Table 7. Socioeconomic data of the most likely cluster within the Nueces County

Socioeconomic	All block groups	The block groups in the most like cluster detected by					
		MLF	NGG	SA	FlexScan	Elliptic SaTScan	Circular SaTScan
Median house income ($)	35,959	31,167	30,469	33,521	26,427	28,419	30,580
Poverty rate (%)	18	21	19	19	24	26	23
Median house built year	1967	1958	1919	1963	1953	1957	1959
Median house value ($)	70,100	58,857	63,074	63,648	49,363	56,033	58,048

5.4. Conclusion

There is an important difference among the performance of traditional SaTScan, FlexScan, SA, and the two algorithms (MLF and NGG) introduced in this chapter. Kulldorff's method tries to search the maximum likelihood ratio using a predefined geometrical shape (circle or ellipse) while the FlexScan method searches for the nearest maximum. For most circular-shape clusters, the spatial scan statistic method will promise fast and efficient cluster detection in many applications. That is why this method is popular in providing an initial analysis for most cluster studies. The two new algorithms make it easy to find out the exact location and boundaries of clusters with arbitrary shapes. Moreover, by adopting the idea of global-optimization strategies, the two new algorithms reduce the effects of the local maximum problem by searching for the global maximum of the likelihood ratios at each step.

Comparing the detected clusters from the two new algorithms and those from SaTScan, FlexScan, and SA, we found the performance of the neighbor-expanding method has been significantly improved in the cluster with arbitrary shapes. However, the computation time of the NGG algorithm was much longer than that of the MLF algorithm. This might be caused by the no-constraint rule when the NGG selects the seed to detect the next level cluster in the search process. Without any penalty on the shape of the result, the NGG allows more detected clusters than the MLF and SA. One possible solution for this problem is to set the degree allowing irregular shape in the detected cluster according to some appropriate criteria, minimizing the occurrence of false clusters. Or the post-process of the entire detected result should be conducted after cluster analysis to remove the highly irregular ones. But this solution will require more detection time and expert knowledge in selecting an appropriate threshold.

One of the most critical components of environment epidemiology is to estimate the associations between human exposures and health outcomes (Nuckols et al. 2004; Ozkaynak et al. 2008). In order to further understand the etiology of a disease, it is necessary to explore the proximity, frequency, and magnitude of potential environmental hazards and their effects to humans. Obviously, this cluster analysis will help understand the geographic distribution of murine typhus in Texas. From this cluster analysis, it can be concluded that the most likely cluster of murine typhus is mostly distributed in warm and humid areas – notably eastern Nueces County along coastal Texas. Moreover, at the census block group level, most of the detected clusters (> 80% or 90%) are in high population density areas (population > 1000 per square kilometer) with lower household incomes and home values. These findings prove that the distribution of murine typhus is controlled by both environmental and socio-economic factors.

The choice of scale/resolution in cluster analysis deserves some attention. In most of case studies, it is preferred to choose a resolution small enough to represent most disease distribution in a relatively homogeneous area. Furthermore, the spatial aggregation of areal data may change the pattern of disease and bring some difficulty in validating the results due to effects of the modifiable areal unit problem (MAUP). A possible solution to this problem involves performing the cluster analyses at different scales of area units to estimate the effects of MAUP and this issue will be addressed in future research. If possible, it would be much better to conduct an analysis of scale effect before conducting a cluster analysis. The choice of scale/resolution for specific cases or specific diseases at different regions should be treated differently. Although there is no specific rule to follow, users of the algorithms should be very familiar with the characteristics of the disease in question as well as the study area before the cluster detection is conducted.

CHAPTER 6

VISUAL EXPLORATION OF MULTIVARIATE ENVIRONMENTAL HEALTH DATA: DETECTION OF LINE CLUSTERS

6.1. Introduction

One critical component of environmental epidemiology involves estimating the associations between human exposures to environmental hazards and health outcomes (Nuckols et al. 2004; Ozkaynak et al. 2008). To further understand the etiology of a disease, it is necessary to explore how human exposure to different environmental hazards would impact human health. Different environmental hazards can be represented by different variables. Therefore, it is quite critical to identify the clustering of multiple variables in relation to a disease and this topic will be the focus of this chapter.

In an attempt to gain insights into complex multivariable data, a number of interesting statistical techniques have been proposed. These techniques include local indicators of spatial association (LISA) (Getis and Ord 1996), spatial autocorrelation analysis (Goodchild 1985; Griffith and Arnrhein 1991), multidimensional scaling (Cliff et al. 1981; Cliff et al. 1995), and trend-surface analysis (Unwin 1975; Gesler 1986). Many of these spatial analytical techniques, combined with GIS tools, have been successfully used in epidemiology and health research (Moore and Carpenter 1999).

With the advance in computer technology, graphic and computational methodologies have been developed with the purpose of exploring complex structural relationships among spatial data. Most of early graphic tools, such as scatter diagrams and bar charts, have shown their advantage in exploring one or two variables in two-dimensional Euclidean space (Chou, Lin and Yeh 1999). However, these traditional tools are very difficult to display or to use in exploring multiple variables or data with more than three dimensions. As an alternative technique, visualization plays a critical role in revealing hidden information about the structures or patterns of spatial data.

Among the visualization methods, parallel coordinate plot offers an excellent tool for exploring the complex interrelationships among multiple variables. Proposed by Inselberg (1985) as a tool for computational geometry (Miller and Wegman 1991), this technique has been applied to data exploration (Bolorforoush and Wegman 1985), cluster identification (Chou, Lin, and Yeh 1999), and data visualization (Klemz and Dunne 2000). Labeling each axis as one variable, the parallel coordinate plot connects points on the adjacent parallel axes using straight lines. This technique overcomes the weakness of traditional Cartesian plots since there is no limitation on the number of dimensionality while the Cartesian plots can only represent the data on two orthogonal coordinate axes (Miller and Wegman 1991). However, the parallel coordinate system analysis has one major limitation: it does not offer any type of statistic measure as regressive techniques to identify or assess the relationships among variables. This problem is more obvious when the size of data becomes very large with overplotting occurring within the scatterplots. Klemz and Dunne (2000) suggested more traditional conclusive research techniques should be performed after an analysis using the parallel coordinate visual. Miller and Wegman (1991) constructed line densities for parallel coordinate plots to display raw data with a density plot for applications involving large datasets. Johansson et al. (2005)

proposed a high-precision texture method to represent the cluster characteristics of parallel coordinates displays.

In this chapter, a new line cluster detection technique was developed and combined with the visual exploration technique, parallel coordinate plots, to reveal the spatial relationship between two variables in multivariable datasets. The line cluster detection technique is based on Kulldorff's spatial scan statistic to investigate the structure of clustered line segments in parallel coordinate plots. The major contribution of this study lies in developing a new method for displaying and revealing line structures within high density parallel coordinates plots. Building upon existing techniques, this study brings together new visual exploration technique and spatial statistical analysis. The potential application of this methodology is to detect the impact of environmental hazards on the identified disease as well as to indicate the potential exposure distance for each of these individual hazards.

6.2. Methodology

The methodology section includes two parts. The first part describes how to generate a parallel coordinate plot to display the relationship among multiple variables. The second part presents the statistical method used to interpret the different types of line cluster patterns.

6.2.1. Parallel Coordinate Plots

The parallel coordinate plot method was designed to visualize N-dimensional data in a R^2 Euclidean space (Inselberg 1985). Historically, the statistical relation between dependent and independent variables was interpreted by xy or xyz scatter plots (Figure 24). This commonly-used method becomes impractical if the data has more than three dimensions. In the parallel coordinate plot, the xy-Cartesian coordinates is replaced by N parallel axes labeled as X1, X2, and X3…. These axes are equidistant and perpendicular to the x-axes, all having the same orientation as the y-axes.

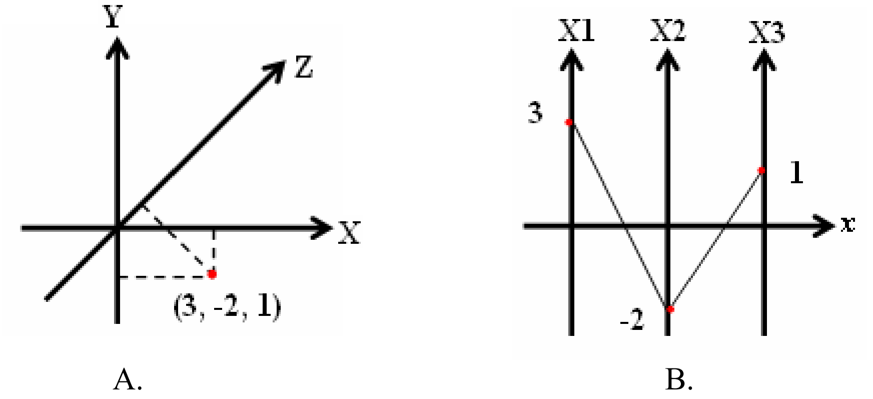


Figure 24. Conversion of Cartesian coordinate system to a parallel coordinate system (Klemz and Dunne 2000).

In the traditional Cartesian coordinates, one point in the N-dimension space has coordinates (A_1, A_2, A_3,…, A_N). As a simple example, Figure 24A illustrates a point (3, -2, 1) in the three-dimensional xyz-Cartesian system. When the dimension number of data increases, it is difficult to display them in the same plane. Using the parallel coordinate system, the coordinates convert into the vertices (i, Ai) in the parallel coordinates on the X1, X2, X3,…, Xn while i = 1, 2, …, N. The sample example is illustrated in Figure 24B. It is obvious that the parallel coordinate plots can display N-dimensional Cartesian points using simple line segments.

In this study, parallel coordinate plot was adopted to represent the multivariable data. Unlike the parallel coordinate plot in which each observation is represented by an unbroken series of line segments, only a line segment will be ploted when both of the two consecutive variables have non-null values. As shown in the following picture, there are segments between AB, DE (red solid line in Figure 25) while no line segment exists between BC and CD since C has a null value in my source data. The blue dash line is used in the Figure 26 to represent the virtual segments.

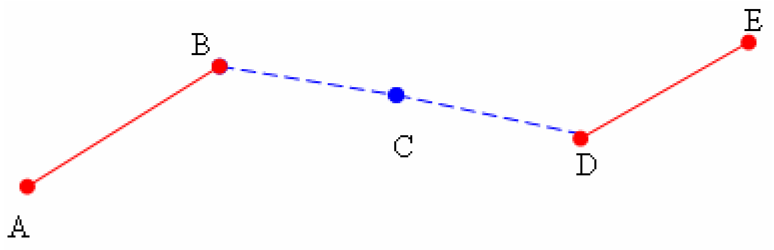


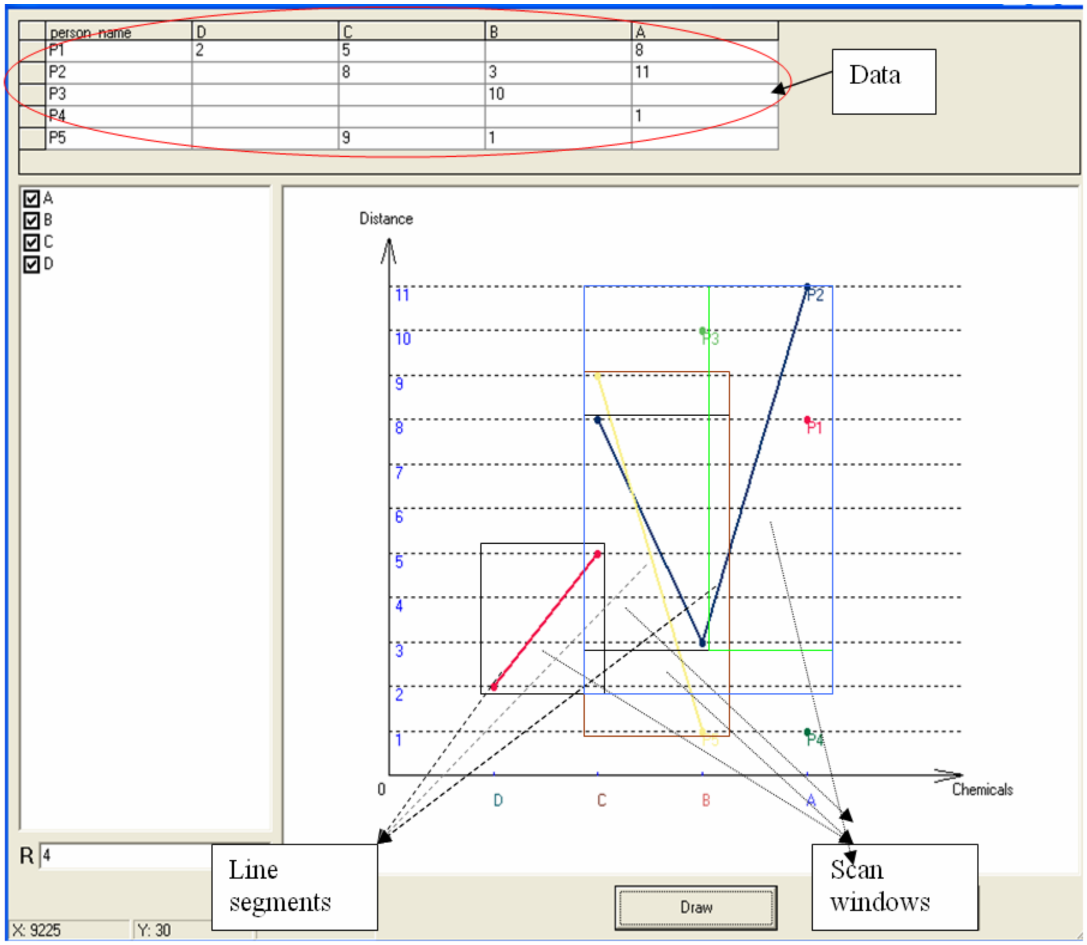
Figure 25. Line segments used to represent the non-null variable and null variable.

After visualizing the multivariable in the dataset, the proposed cluster technique was applied to the line segments to identify the line pattern within these multiple variables.

6.2.2. Representing Line Cluster Characteristics Using the Spatial Scan Statistic

A modified spatial scan statistic was applied to analyze the visualization result, through the following three major steps:

First, a set of various-size rectangles was generated as scan windows (Figure 26) to cover the whole set of line segments. These rectangles have their widths measurements which are either equal to or multiples of the distance of two consecutive variables along the x-axis and various vertical lengths. The number of line segments in any rectangular scan window were counted and used in the next step.



person_name	D	C	B	A
P1	2	5		8
P2		8	3	11
P3			10	
P4				1
P5		9	1	

☑ A
☑ B
☑ C
☑ D

R 4

X: 9225 Y: 30

Figure 26. Procedures of the proposed method.

Secondly, the likelihoods of each rectangular scan window were calculated based on the number of line segments. Based on the theory from Kulldorff's spatial scan statistic, the likelihood ratios were calculated based on two particular models: the Bernoulli model and the Poisson model. The following notations and equation were used in this chapter. Let Z be the set of rectangular scan windows, μ(G) represent the number of all ***possibl***e line segments (***red solid line segments and blue dash line segments***). For instance, in the above Figure 26, there are four line segments: AB, BC, CD, and DE. Let **nG** denote the total number of observed line segments (***red solid line segments***), **μ(z)** stands for the number of all possible line segments in **z**th scan window**,** and $\mathbf{n_z}$ be the observed number in **z**th scan window. Moreover, ***p*** is defined as the probability that an incidence falls in the **z**th scan window, ***q*** as the probability that an incidence falls in the rest of the study area, and ***p***, ***q*** as the numbers between 0 and 1. Now it can be tested whether the specific scan window **z** is a cluster. The null hypothesis is H_1: $\boldsymbol{p}$=$\boldsymbol{q}$, **z** ϵ **Z** and the alternative hypothesis is H_0: $\boldsymbol{p}$>$\boldsymbol{q}$, **z** ϵ **Z.** The former represents that the probability is constant for all scan windows while the latter stands for the probability that an incidence in scan window z is different from the probability that any incidence outside the scan window z.

For scan window **z**, the likelihood function based on the Bernoulli model can be expressed as the follow formula:

$$L(z) = \sup_{p>q} L(z,p,q) = (p)^{n_z} \times (1-p)^{\mu(z)-n_z} \times q^{nG-n_z} \times (1-q)^{(\mu(G)-\mu(z)-(nG-n_z))} \qquad (1)$$

The observed likelihood function for scan window z can be given by

$$L(z) = \begin{cases} \sup_{p>q} L(z,p,q) = (\dfrac{n_z}{\mu(z)})^{n_z} \times (1-\dfrac{n_z}{\mu(z)})^{\mu(z)-n_z} \times & \\ (\dfrac{nG-n_z}{\mu(G)-\mu(z)})^{nG-n_z} \times (1-\dfrac{nG-n_z}{\mu(G)-\mu(z)})^{(\mu\mu(G-\mu(z)-(nG-n_z))} & if\ (\dfrac{n_z}{\mu(z)}) > (\dfrac{nG-n_z}{\mu(G)-\mu(z)}) \\ or & \\ (\dfrac{nG}{\mu(G)})^{nG} \times (\dfrac{\mu(G)-nG}{\mu(G)})^{\mu(G)-nG} & otherwise \end{cases}$$

(2)

Thus the expected likelihood function are:

$$L_0 = \sup_{p=q} L(Z,p,q) = (\frac{nG}{\mu(G)})^{nG} \times (\frac{\mu(G)-nG}{\mu(G)})^{\mu(G)-nG} \qquad (3)$$

Therefore the likelihood ratio λ(z) can be obtained as a quotient by dividing the observed likelihood by the expected likelihood:

$$\lambda(z) = \begin{cases} \dfrac{L(z)}{L_0} = \dfrac{\sup\limits_{p>.q} L(z,p,q)}{\sup\limits_{p=q} L(Z,p,q)} & if\ (\dfrac{n_z}{\mu(z)}) > (\dfrac{nG - n_z}{\mu(G) - \mu(z)}) \\ or & \\ 1 & otherwise \end{cases} \tag{4}$$

Another likelihood ratio test is based on the Poisson model as presented below:

$$\lambda(z) = \begin{cases} \dfrac{L(z)}{L_0} = \dfrac{(\dfrac{n_z}{\mu(z)})^{n_z} \times (\dfrac{nG - n_z}{\mu(G) - \mu(z)})^{nG - n_z}}{(\dfrac{nG}{\mu(G)})^{nG_z}} & if\ (\dfrac{n_z}{\mu(z)}) > (\dfrac{nG - n_z}{\mu(G) - \mu(z)}) \\ or \quad 1 & otherwise \end{cases} \tag{5}$$

After calculating the likelihood ratios for all scan windows using the set of formulas based on either one of the two models, the most likely cluster can now be identified by searching for the scan window with the maximum likelihood ratio.

The last step is to test the statistical significance. The significant of likelihood ratio ***λ*****(z)** could be examined using the Monte Carlo simulation. Given the total number of observed line segments, the Monte Carlo simulation will distribute them randomly on the map under the null hypothesis. A larger number, ***N,*** of simulations, for example, 9999 times, were performed. For each simulation, the likelihood ratio ***λ*** (**z**) of scan window **z** were calculated in the same way but using the simulated data. Then the number of simulations, $\boldsymbol{n}$, was calculated in which the simulated $\boldsymbol{\lambda}(\mathbf{z})$ was larger than or equal to the $\boldsymbol{\lambda}(\mathbf{z})$ obtained from the real data. The p-value can be estimated by $\boldsymbol{n}/(\boldsymbol{N+1})$.

6.3. Results and Discussion

6.3.1. An Initial Bi-variable Analysis

A regression analysis between every two environmental variables was analyzed using the matrix scatter diagram and the results show in Figure 27. The regression equation is shown with this symmetrical bi-variable matrix scatter diagram. Most of environmental variables are related to other variables and their R^2 (coefficient of determination) is relatively high. The highest R^2 (0.9896) is found between DWTSOLV (average dry weight of solvent) and DWTAHSO (average dry weight of aHsO). The second highest R^2 (0.9608) is found between DWTARSOL (average dry weight of aromatic solvent) and DWTAHSO (average dry weight of aHsO). The least related hazards are DWTMETAL and DWTPAH with having the lowest R^2 (0.5503).

This scatter matrix diagram displays the relationship between each two variables very well by showing both the regression equation and coefficient of determination R^2. However, this is not enough when more variables are involved such as to identify the impact of multiple environmental hazards on the identified disease cases. The relationship interpreted by the scatter matrix diagram only implies the co-occurrence of two variables, without identifying the relationships among these multiple variables. Second, this bi-variable analysis can only explore the relationship between each two variables while in the real world it would be possible to involve multiple dimensional data. The scatter matrix diagram method limits the impact from other variables by only considering two variables one time. In this research, we performed a further analysis using an advanced visualization technique and Kulldorf's spatial scan statistic to detect the spatial patterns for the potential research on the co-impact of multiple variables.

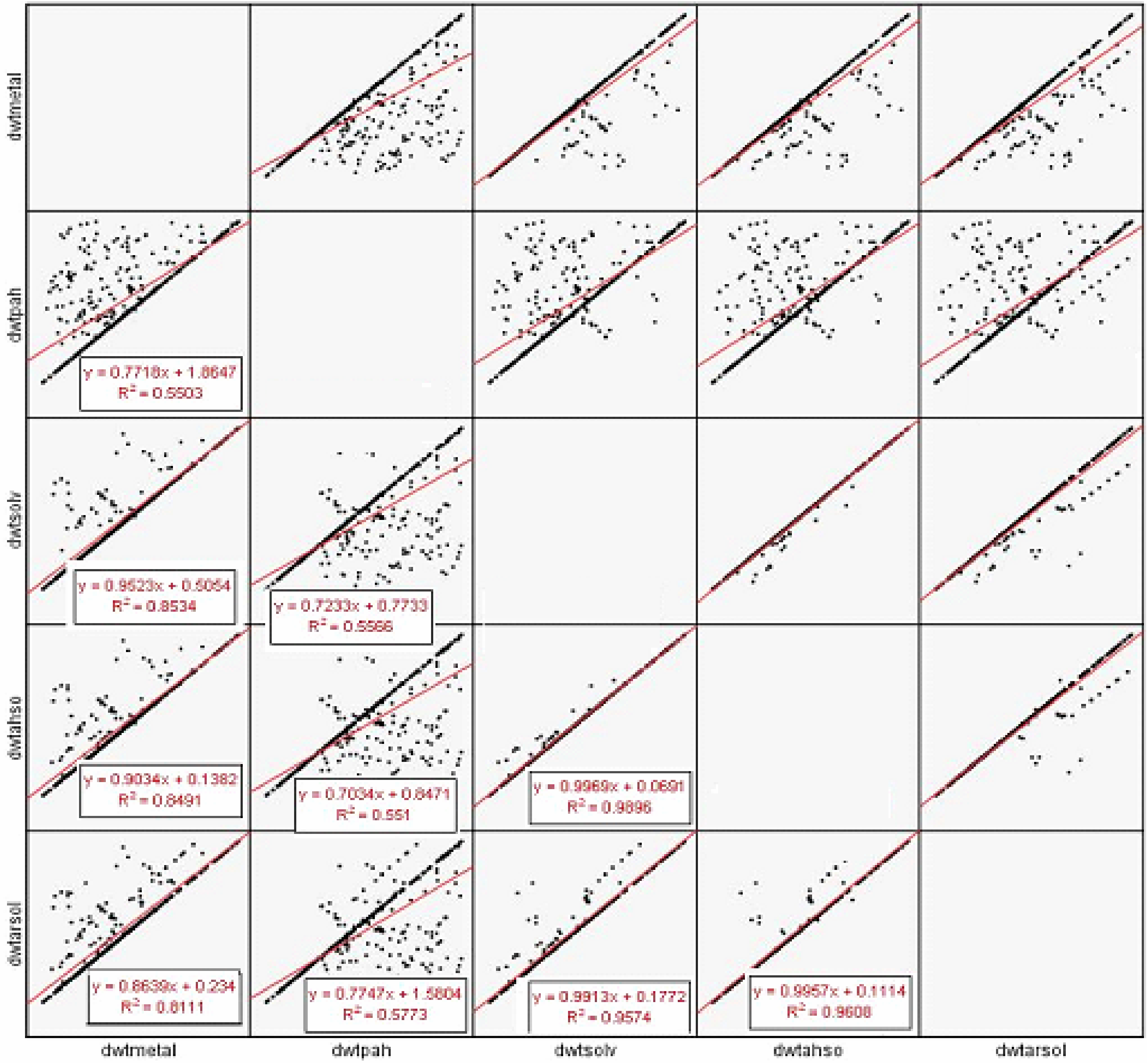


Figure 27. The scatter matrix diagram showing the bi-variable relationship.

6.3.2. Parallel Coordinate Plots

The parallel coordinate plots (PCP) method was adopted in this chapter to investigate the relationship among multiple variables. Figure 28 shows some initial results. In Figure 28, the x-axis is the variables and the y-axis is the distance between each identified case and all variables. In order to visualize the relationship between any two chemicals, all possible combination of every two variables is listed in the x axis. It can be easily found that the high density areas of lines are between variable 3 (average dry weight of solvent) and variable 5 (average dry weight of aromatic solvent) which is also the second highest correlation in the scatter diagram. The lowest line density is found between variable 1 (average dry weight of metal) and variable 2 (average dry weight hydrocarbons) which correlates least as a pair in the matrix scatter diagram. This match further validates the results from both matrix scatter diagram and parallel coordinate plots.

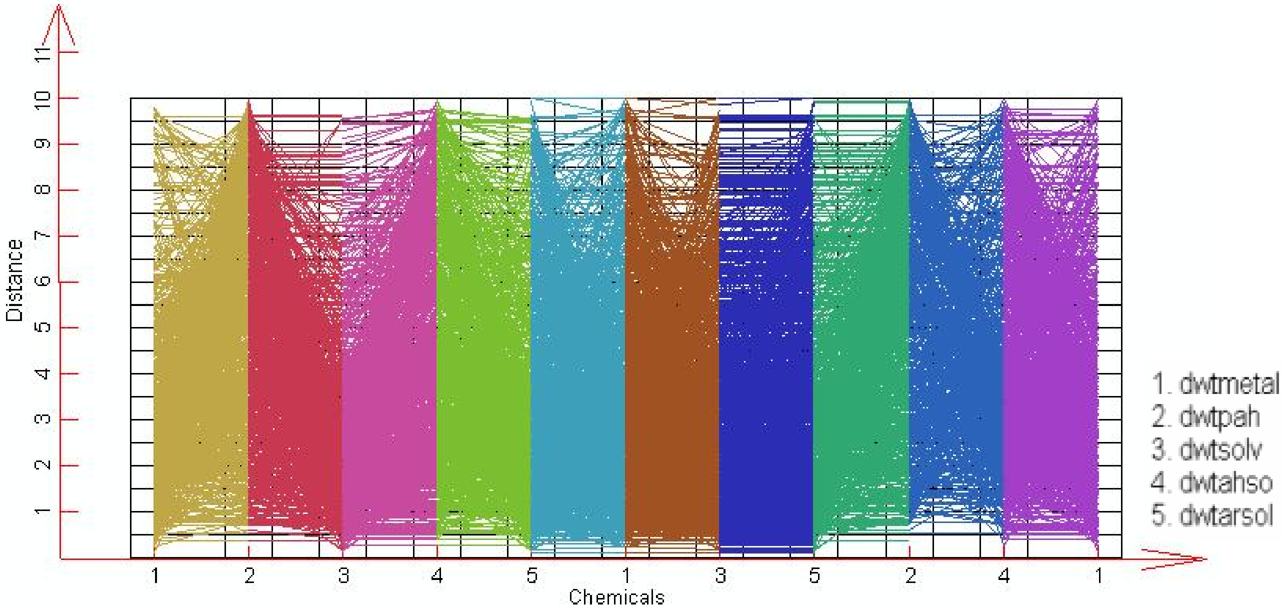


Figure 28. The line segment plots.

Generally, the major contribution of PCP is twofold. First, this technique offers a new insight into the data structure without the extensive training in mathematics or statistics. This convenience offers a simple and practical tool to explore the complex data relationships in high dimensional data. Moreover, this technique helps gain insights into complex structures of multivariate data in large size without knowing the expected relationships among them. However, when there are a large number of cases, the "over-plotting" problem is unavoidable and it is very hard for us to perceive any structures or trends.

In this research, the Kulldorf's spatial scan statistic method was modified and applied for line displaying in the parallel coordinate techniques. Originally used for point data, this modified method still follows the neighbor-expanding idea when searching the cluster pattern of lines (Figure 29) and gradually expanded for each plot. The lines completed inside the search square were counted and calculated using equations (1)-(5). The result is illustrated in Figure 30. The darker the color of the scan window, the higher LLR value this window has. Obviously, the higher LLR (likelihood ratio) is found between variable 3 and variable 5 and the lowest LLR is found between variable 2 and variable 4. This result is very close to that found in the bi-variable analysis using matrix scatter diagram.

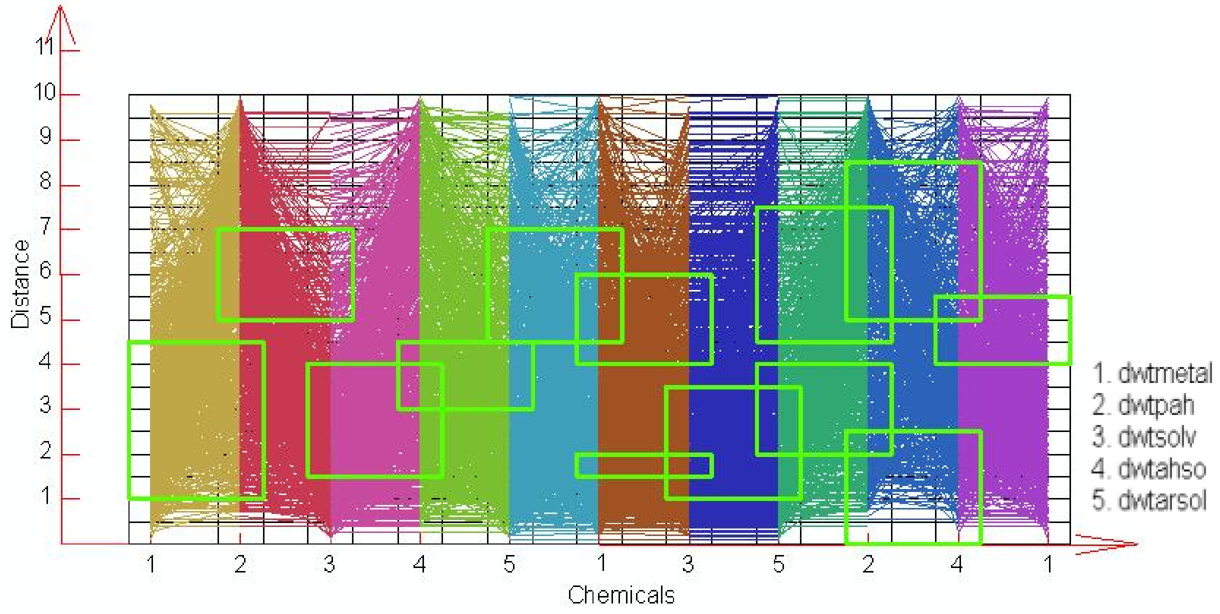


Figure 29. A subset of randomly selected scan windows.

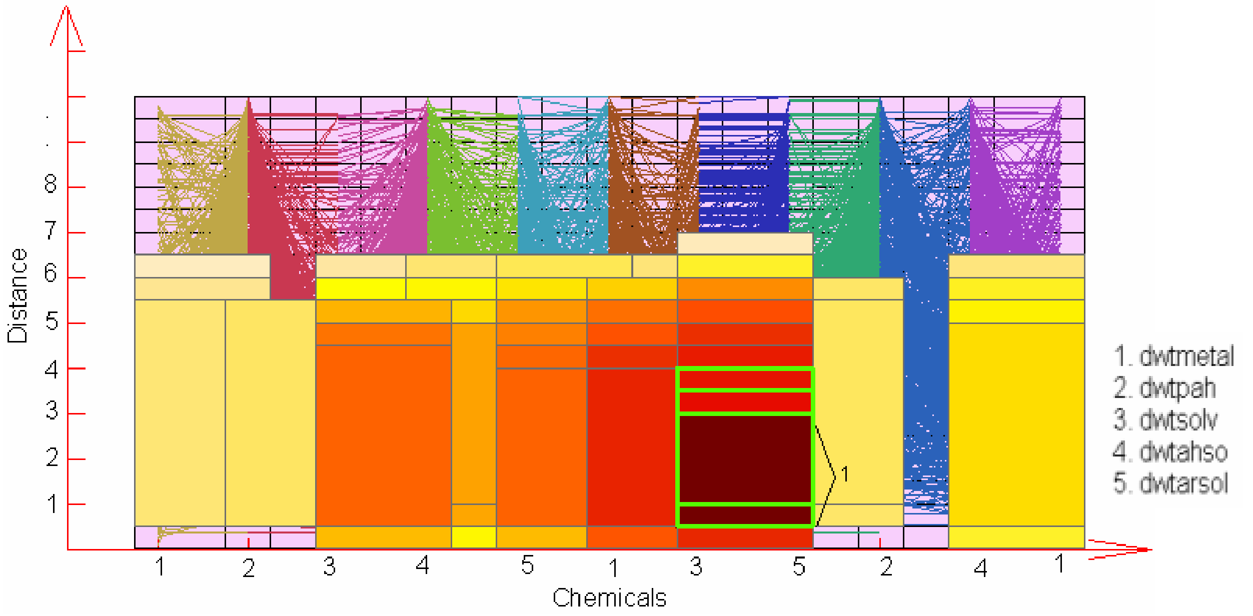


Figure 30. The set of scan windows containing significant line clusters.

Table 8. The first five highest LLRs between DWTSOLV and DWTARSOL

Clusters	Observed #	Expected #	LLR	P-value
7	596	62	829.54	0.001
7	789	122	829.09	0.003
7	688	89	826.69	0.012
7	672	89	793.16	0.037
7	867	159	788.862	0.053

A further analysis was conducted on plots with each combination of every two chemicals as consecutive variables plotter on the diagram. The statistic results show that the maximum LLR is found between variable 3 (average dry weight of solvent) and variable 5 (average dry weight of aromatic solvent) at a distance between 0.5 to 3.0 miles. This means that the highest co-impact distance between variable 3 and variable 5 is between 0.5 to 3.0 miles. Other high LLRs were found at distances of 3.5 miles and 4.0 miles. The highest LLRs, representing the most effective co-impact distance between variable 3 and variable 5, are highlighted in Figure 30 and Table 8. The scale of 1-10 was used to categorize the combination of each two variables; category 7 (combination between variable 3 and variable 5) had the highest LLR and greatest distance combination among these variables. Table 8 shows the highest five LLRs and they are all found in category 7.

Table 9. The first three high LLRs for each category

	Distance Range	Obser-ved #	Expec-ted #	LLR	P-value
1	1.0 to 4.0	232	89	80.32	0.625
	1.0 to 4.5	276	121	73.82	0.649
	1.5 to 4.0	178	62	72.41	0.684
2	1.0 to 4.0	255	89	103.83	0.458
	0.5 to 4.0	306	122	99.12	0.526
	0.5 to 3.5	246	89	94.37	0.593
3	0.5 to 4.5	573	159	329.48	0.106
	1.0 to 4.5	488	121	320.49	0.129
	0.5 to 4.0	486	121	317.63	0.143
4	1.0 to 4.5	453	122	268.93	0.182
	1.0 to 5.0	519	158	263.01	0.194
	1.0 to 4.0	377	89	260.50	0.207
5	0.5 to 4.0	631	121	545.61	0.043
	0.5 to 4.5	721	159	544.45	0.065
	1.0 to 4.5	615	121	518.55	0.089
6	0.5 to 4.0	737	121	735.38	0.023
	0.5 to 3.5	631	89	709.14	0.031
	0.5 to 4.5	815	159	689.49	0.036
7	0.5 to 3.0	596	62	829.54	0.001
	0.5 to 4.0	789	122	829.09	0.003
	0.5 to 3.5	688	89	826.69	0.011
8	1.0 to 4.0	264	89	113.62	0.432
	1.0 to 4.5	306	121	100.66	0.51
	1.0 to 3.5	198	62	94.85	0.586
9	2.0 to 4.5	141	63	35.90	0.782
	2.0 to 4.0	104	40	35.58	0.799
	1.5 to 4.5	179	89	35.49	0.816
10	1.0 to 4.5	422	121	230.84	0.239
	0.5 to 4.5	490	159	226.17	0.251
	1.0 to 5.0	490	159	226.17	0.251

Note: 1: lines between metal and hydrocarbons; 2: lines between hydrocarbons and solvent;
3: lines between solvent and aHsO; 4: lines between aHsO and aromatic solvent;
5: lines between aromatic solvent and metal; 6: lines between metal and solvent;
7: lines between solvent and aromatic solvent; 8: lines between aromatic solvent and hydrocarbons;
9: lines between hydrocarbons and aHsO; 10: lines between aHsO and metal.

In order to find out the statistical significance of this cluster pattern analysis, the highest LLR for each two variables are listed in Table 9. It is easily noted that the most common distance range for all the categories falls between 1.0 to 4.5 miles. This means that within this distance, the highest co-impact from each pair of variables will occur. The high LLRs were found in category 7 (relationship between solvent and aromatic solvent) and category 6 (relationship between metal and solvent). The low LLRs were found in category 9 (relationship between hydrocarbons and aHsO) and category 1 (metal and hydrocarbons). Many researchers have found that the relationship between the oral cleft and the exposure to solvents during pregnancy is significant (Holmberg et al. 1982; Laumon et al. 1996; Chevrier et al. 2006). Co-existence between the solvents will actually reinforce their impact; this is why the highest impact is found between solvent and aromatic solvent in this range of distance.

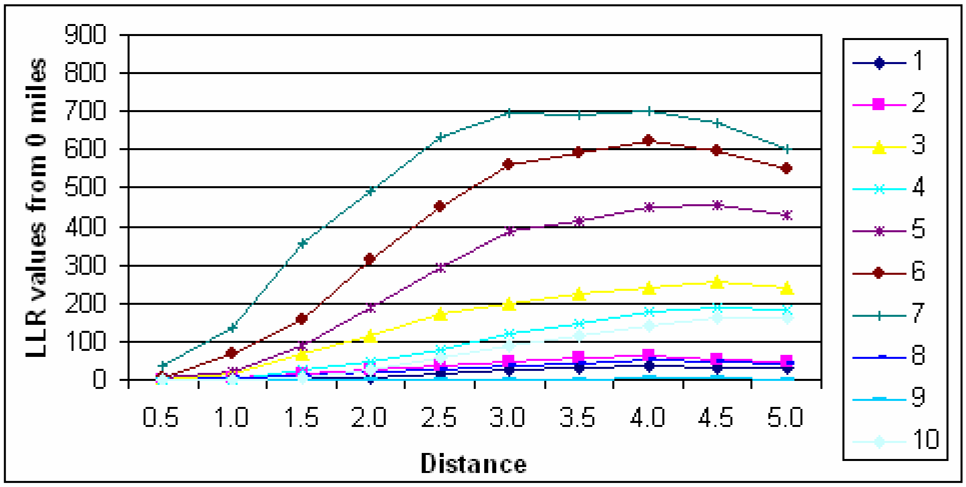
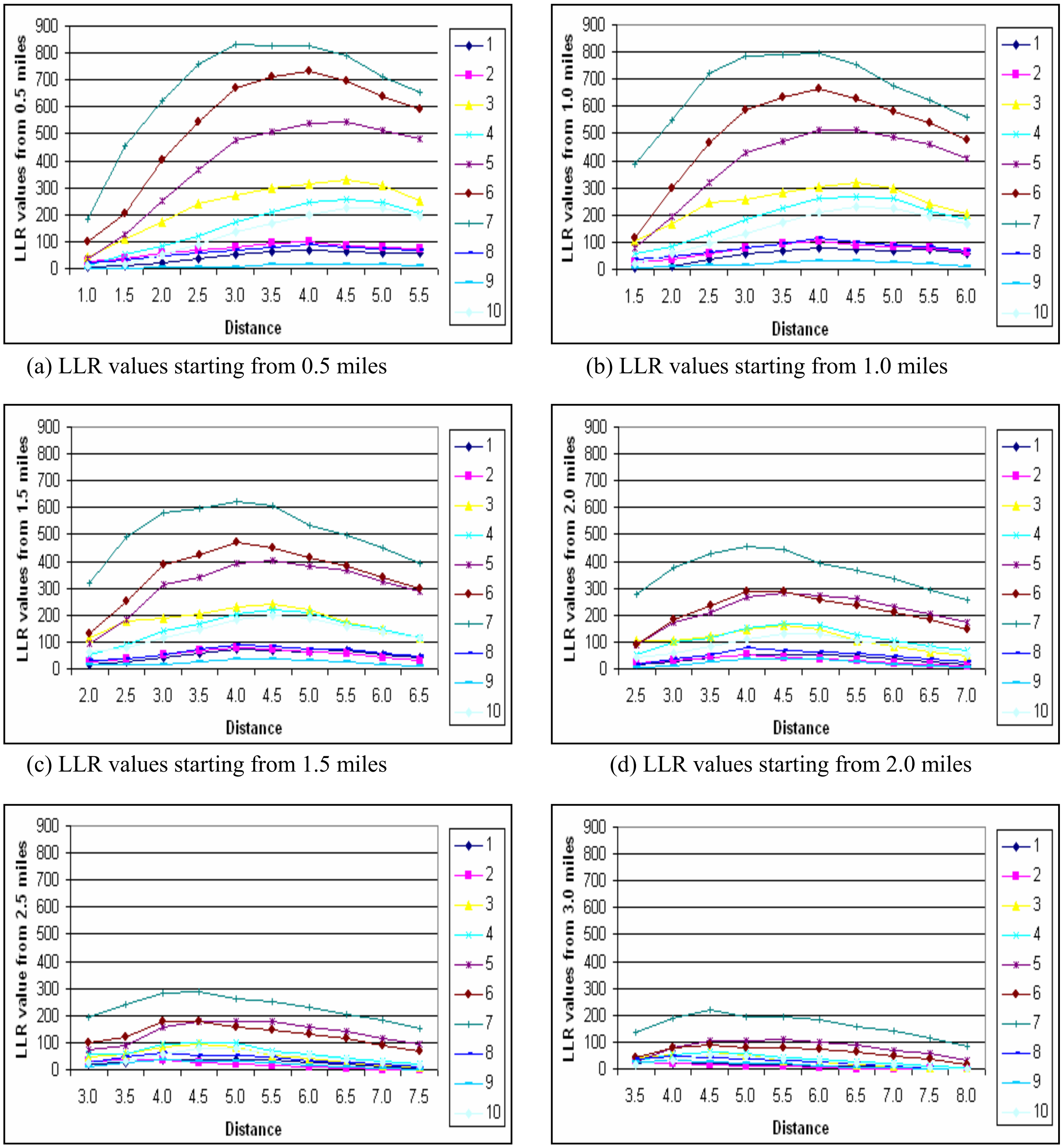


Figure 31. The LLR trend for each category at increasing distance.
Note: The category is the same as table 9; the starting distance is 0 miles.

In order to identify the most significant distance and potential effects of co-impact distance among variable pairs, the LLRs for all categories were examined using scanning windows starting at 0.5 miles and ending at 5.0 miles. The results are shown in Figure 32. Obviously, the highest LLRs found in category 7 remain the same as the one shown in Figure 31. For most categories, the highest LLR is always found at 4 miles. These results indicate that the four-mile mark might be the distance having the highest influence from these multiple variables (chemicals).



(a) LLR values starting from 0.5 miles

(b) LLR values starting from 1.0 miles

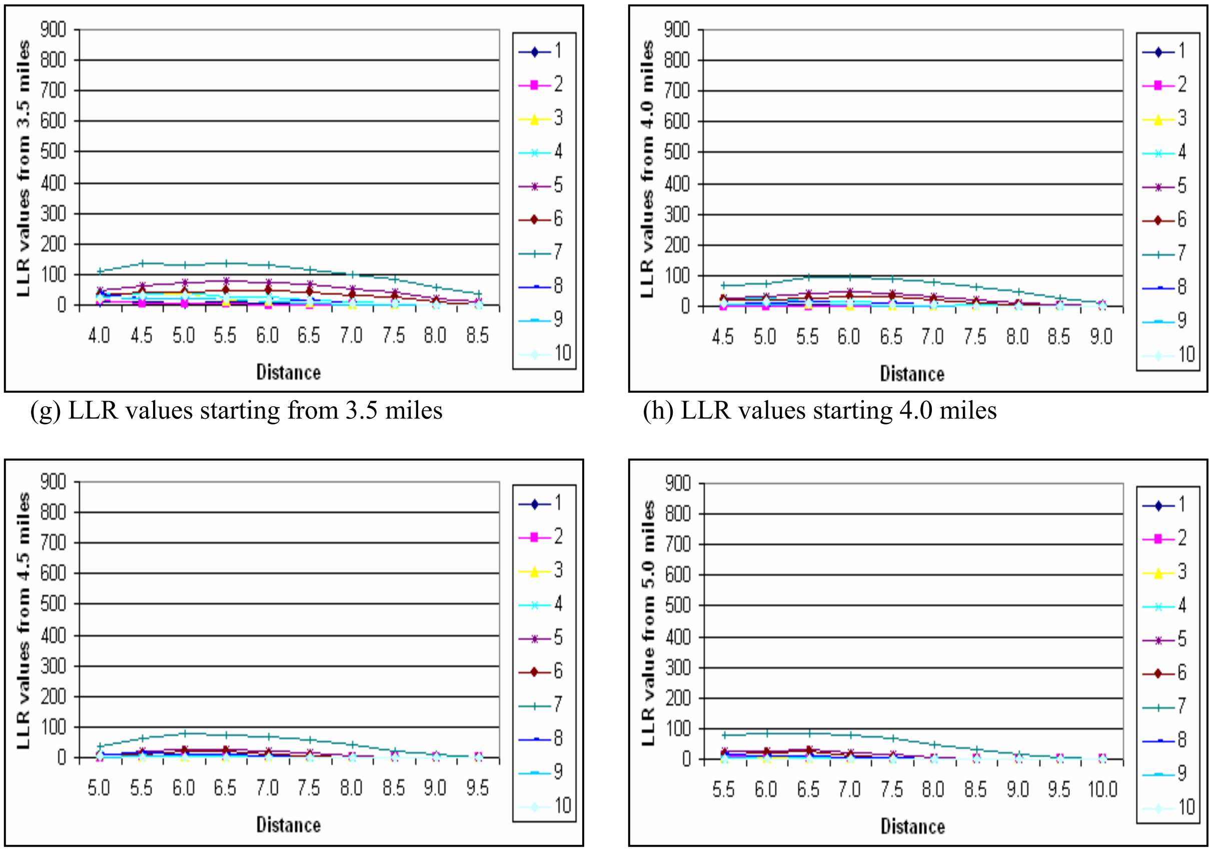
(c) LLR values starting from 1.5 miles

(d) LLR values starting from 2.0 miles

(e) LLR values starting from 2.5 miles

(f) LLR values starting from 3.0 miles

Figure 32. The LLR trend for each category at increasing distance starting from 0.5 miles.
Note: The categorizing range is the same as the one shown in table 9.



(g) LLR values starting from 3.5 miles

(h) LLR values starting 4.0 miles

(i) LLR values starting from 4.5 miles

(j) LLR values starting from 5.0 miles

Figure 32. Continued

The LLR values from 1.0 to 5.0 at 0.5 increments for all categories are shown in Figure 32. For all distance intervals, the high LLRs are still found in category 7 with the highest one found in the distance from 0.5 to 3.0 miles. This result is consistent to the result we found in Table 8. Another interesting thing in these figures is that the LLR values in Figure 32 are not the highest LLR as expected. In the hypothesis, the highest impact would come from the closest variable/chemical, assuming the impact would be highest if two chemicals were very close to each other. However, most of the LLRs continue to increase even when the scanning windows start from 0.5 mile instead of from 0 mile. Moreover, they increase as the searching scan window becomes larger than 3.0

miles and only decreases from 4.0 miles. This suggests that the highest impact from multiple variables is not always found in the closest proximity of these multiple variables. The possible reason might be the insufficient cases living close to the hazardous materials within 0.5 miles. In this study, the obvious threshold is found between 3.0 to 4.0 miles which appears to be the most dangerous place for the identified disease case in this study.

6.4. Conclusion

In this chapter, the spatial scan statistic method was adopted to reveal the structure in line clusters visualized by parallel coordinate plots. The method is the same as the one discussed in Chapter 5 by expanding neighbors as scan windows. An obvious result could be derived from the analysis: through testing the disease samples we have, it is very useful to detect the line cluster pattern to find out the relationship among multiple variables and the LLR values also could be used to test the significance of line cluster pattern. This method could be applied to identify the potential co-impact from multiple environment hazards and the most dangerous place with the highest impact from these hazards.

This research is based on two existing techniques—parallel coordinates plots and neighbor-expanding techniques—which is presented in Chapter 5. Our research results can be further improved in several aspects. First, it would be interesting to study whether advanced pattern analysis methods could be used to further reveal the “hidden” cluster properties. The spatial scan statistic method was originally developed for analyzing spatial point data instead of line distribution and it is not conclusive that this spatial scan statistic method is appropriate when applied to line data. It is still necessary to compare and evaluate more sophisticated line cluster detection methods in more case studies before assuming the validity of spatial scan statistic method in linear patterns.

CHAPTER 7

SUMMARY AND FUTURE WORK

This chapter summarizes the study by offering a discussion of the analysis along with a research summary. Both the limitations and potential contributions of this research are also discussed. After the research summary, the closing section provides a discussion of the directions of future work that could be built upon this research.

7.1. Research Summary

The general goal of this doctoral research was to develop and examine a new method to automatically detect cluster patterns of arbitrary shapes in health data. The first research objective was to develop two algorithms for the cluster detection. Among the automatic detection methods, the spatial scan statistic model is one of the most-commonly used methods. The major problem of this method lies in the usage of regularly-shaped scanning windows which limits its application to find geographic regions of general shapes where a cluster of disease concentrations may exist. A new neighbor-expanding approach is developed based on a maximum-likelihood-first algorithm and non-greedy growth algorithm. In addition, this method is evaluated using a dataset of murine typhus in Texas. The detailed algorithms were described in Chapter 5. By searching any arbitrarily shaped scan window, this neighbor-expanding approach was applied to detect the arbitrarily shaped cluster pattern of murine typhus in south Texas.

Comparing the detection results from the traditional spatial scan statistic, maximum-like-first method, and non-greedy growth algorithm, the maximum-like-first method performs better than the spatial scan statistic. The shape of detected objects affects the performance of the proposed non-greedy method; the computing time of the non-greedy algorithm is much longer than the other two methods. Unlike the other two methods, the non-greedy method allows more clusters detected with no constraints on the shape of detected shapes. All three methods imply that the detected most likely clusters are found in an area with high population density (with population density > 1,000 per square kilometer) as well as warm and humid environment along coastal Texas.

Another research objective in this study was to ascertain the feasibility to apply the neighbor-expanding method, with the aid of visual exploration, to detect the line cluster patterns involving multiple variable analyses. Using the distance between the locations of identified disease incidences and the location of potential environment hazards, this research explores the potential of this proposed method to investigate the simultaneous impact of exposures to two containments in nearby environmental hazards for disease analysis.

7.2. Limitations

There are several limitations in this doctoral research. One of the main limitations is the time-efficiency of neighbor-expanding methods since it tries to search all neighboring areas. Technically, this research is built upon the most popular cluster detection method, the spatial scan statistic method. Although the neighbor-expanding method improves the spatial scan statistic method for cluster analysis involving clusters of arbitrary shapes and two variables, it still needs much improvement in time efficiency and accuracy. Therefore, these algorithms are not practical for large areas with huge volumes of data. Moreover, a detailed accuracy assessment was not undertaken for the line cluster analysis of this dissertation study since the spatial location was not available for each disease case, but this is still a necessary step for both methods in disease analysis. The accuracy is relatively high for point data; the linear data need more test and validation in the future.

For the point disease data, the major limitation was that the locations of disease cases were the only useful information collected in the research area. The major transfer entities are small mammals such as domestic cats, mice, opossums and the fleas associated with them. It is very hard, if not impossible, to consistently determine the locations of these small mammals. Moreover, the movement of small mammals is very general and very hard to trace over a large geographic area. Because of the lack of information, the spatial distribution of mammals was assumed to be determined mainly by environmental factors. For this reason, only the population density was selected and studied for the impact analysis of environment.

Another limitation was insufficient data for more fully accomplishing the second research objective. Based on the proposed method, further research on the etiology of disease will be a very interesting topic. However, the only available data was the distance between the locations of various environmental hazards and the locations of the cases. Although the disease may be caused by many factors, such as a patient's age or a patient's residential location, it was not possible to extend this research to a detailed analysis due to the lack of data. The correlation between the environment and individual health reports should be analyzed first before we can perform significant evaluations concerning environment hazards. Further, some patients may have lived in several locations before they became sick. Due to the lack of this information, it was only possible to conduct the research as presented in this dissertation.

7.3. Future Work

The methodology developed and reported in this research has many potential applications. Based on the results derived from this research, there are many potential studies that can be further explored. The algorithm and methodology development was the major focus of this dissertation. The major contribution is to determine how to improve the existing cluster detection methods. The inclusion of both pathology and environmental exposure was important for this research. Inclusion of other exposure assessments would help us understand the relationship between disease and environmental exposures. In the future, additional information about individual cases, such as a patient's age, race, and health history would need to be involved in the cluster detection and pattern analysis.

Another future project would be the improvement of the algorithms. In order to apply it to a more general study, the algorithms still need more calibration and optimization with respect to both efficiency and accuracy. Both the maximum-likelihood-first algorithm and non-greedy algorithm are step-by-step algorithms which make the whole process very slow. In the future, it would be much efficient if the simultaneous detections are introduced to detect several cluster centers at the same time or preselected by the detectors. This simultaneous detection would need higher computational power to perform a quick search. Improved speed will make this technique more practical for cluster detection in large study areas with high volumes of data.

The multiple-year and cross-site tests would be an interesting topic in the future. More data would be collected to support tests covering multiple years and multiple sites. With other ancillary data, such as demographic data and environmental data, it is feasible to develop a sophisticated model to predict a potential cluster pattern. If there are a certain number of cases reported, this model could be used to predict a possible "hot spot" and provide timely warning alerting public health officials.

The development of more functions to calibrate the ambiguity and uncertainty associated with cluster detection and disease data is a fascinating topic. Most cluster detecting tests lack an easy and accessible technique for validating a model. A more efficient and accessible technique would not only allow a better understanding of cluster patterns, but also provide a certain confidence in offering the result to customers. Each of these assessments would provide valuable insight into the influence of data sources and detection methods related to cluster analyses. This type of useful information would likely have a significant influence on the design of cluster detection algorithms and the collection of ancillary data for cluster detection.

REFERENCES

Alexander, F. E. 1999. Clusters and clustering of childhood cancer: A review. *European Journal of Epidemiology* 15: 847-852.

Alm, S. E. 1997. On the distributions of scan statisitcs of a two-dimensional poisson process. *Advances in Applied Probability* 29: 1-18.

Anderson, N. H., and D. M. Titterington. 1997. Some methods for investigating spatial clustering, with epidemiological application. *Journal of the Royal Statistical Society Series* A 160: 87-105.

Andes, N., and J. E. Davis. 1995. Linking public health data using geographic information system techniques: Alaskan community characteristics and infant mortality. *Statistics in Medicine* 14: 481-490.

Anselin, L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis* 27: 93-115.

Assuncao, R., and E. Reis. 1999. A new proposal to adjust Moran's *I* for population density. *Statistical in Medicine* 18: 2147–2162.

Assuncao, R., M. Costa, A. Tavares, and S. Ferreira. 2006. Fast detection of arbitrarily shaped disease clusters. *Statistical in Medicine* 25: 723-742.

Azad, A. F. 1990. Epidemiology of Murine Typhus. *Annual Review of Entomology* 35: 553-569.

Bachmann, M. O. 2003. When is a cluster of disease really a cluster? *Occupational Medicine* 53: 157-158.

Bailey, T. C. 2001. Spatial statistical methods in health. *Cad Saude Publica* 17: 1083-1098.

Baptiste, M. S., R. Rothenberg, P. C. Nasca, D. T. Janerich, C. D. Stutzman, K. Rimawi, W. O'Brien, and J. Matuszek. 1984. Health effects associated with exposure to radioactively contaminated gold rings. *Journal of the American Academy of Dermatology* 10: 1019-1023.

Beyers, N., R. P. Gie, and H. L. Zietsman. 1996. The use of a geographical information system (GIS) to evaluate the distribution of tuberculosis in a high-incidence community. *South African Medican Journal* 86: 40-44.

Bennett, R. J., and R. P. Haining. 1985. Spatial structure and spatial interaction: Modelling approaches to the statistical analysis of geographical data. *Journal of the Royal Statistical Society. Series A (General)* 148, no. 1: 1-36.

Besag, J., J York, and A Mollié. 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, no. 1: 1-20.

Besag, J., and J. Newell. 1991. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society A* 154: 143-155.

Bithell, J. F. 1995. The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine* 14, no. 21-22: 2309-22.

Bithell, J. F. 1999. Disease mapping using the relative risk function estimated from areal data. In *Disease Mapping and Risk Assessment for Public Health*. A.B. Lawson, A. Biggeri, D. Bohning, E. Lesaffre, J.-F. Viel, and R. Bertollini (Eds.) pp. 247-55. New York: John Wiley & Sons.

Blum, H. F. 1948. Sunlight as a causal factor in cancer of the skin of man. *Journal of the National Cancer Institute* 9: 247-258.

Bootstrom, A., M. S. Beier, J. A. Macaluso, K. R. Macaluso, D. Sprenger, J. Hayes, S. Radulovic, and A. F. Azad. 2002. Geographic association of rickettsia felis-infected opossums with human murine. *Emerging Infectious Diseases* 8 (6): 549 - 554.

Burton, Ian. 1963. The quantitative revolution and theoretical geography. *The Canadian Geographer* 7: 151-62.

Caldwell, G. G. and C. W. Jr. Heath. 1976. Case clustering in cancer. *Southern Medical Journal* 69: 1598-1602.

Carl, G., and I. Kuhn. 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modeling* 207: 159-170.

CDC. 1990. Guidelines for investigating clusters of health events. *Morbidity and Mortality Weekly Report* 39: 1-23.

Chen J., R. E. Roth, A. T. Naito, E. J. Lengerich, and A. M. MacEachren. 2008. Geovisual analytics to enhance spatial scan statistic interpretation: An analysis of u.s. cervical cancer mortality. *International of Health Geographics* 7: 57-75.

Chevrier, C., B. Danache, M. Bahuau, A. Nelva, C. Herman, C. Francanner, E. Robert-Gnansia, and S. Cordier. 2006. Occupational exposure to organic solvent mixtures during pregnancy and the risk of non-syndromic oral clefts. *Occupational Environmental Medicine* 63: 617-623.

Clark, P. J., and F. C. Evans. 1954. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* 35, no. 4: 445-453.

Clayton, D., and J. Kaldor. 1987. Empirical bayes estimates of age-standardised relative risks for use in disease mapping. *Biometric* 43: 671-81.

Clayton, D., and L. Berardinelli. 1992. Bayesian methods for mapping disease risk. In Geographical and Environmental Epidemiology. P. Ellicott, J. Cuzick, D. English and R. Stern (Eds.). pp. 205-20. Oxford: Oxford University Press.

Conley J., M. Gahegan, and J. Macgill. 2005. A genetic approach to detecting clusters in point data sets. *Geographical Analysis* 37: 286-317.

Cook-Mozaffari, P. J., S. C. Darby, R. Doll, D. Forman, C. Hermon, M. C. Pike, and T. Vincent. 1989. Geographical variation in mortality from leukaemia and other cancers in england and wales in relation to proximity to nuclear installations, 1969-78. *British Journal of Cancer* 59: 476-85.

Cuzick, J., and R. Edwards. 1990. Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society Series B* 52**:** 73-104.

Diggle, P.J. 1990. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society* 153: 349-362.

Diggle, P.J. and B.S. Rowlinson. 1994. A conditional approach to point process modeling of elevated risk. *Journal of the Royal Statistical Society* 157: 433-440.

Diggle, Peter. 2003. *Statistical analysis of spatial point patterns*. New York: Oxford University Press.

Duczmal, L. and R. Assuncao. 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 45 (2): 269-286.

Duczmal L., M. Kulldorff, and L. Huang. 2006. Evaluation of spatial scan statistics for irregularly shaped clusters. *Journal of Computational and Graphical Statistics* 15: 428-442.

Duczmal L., G. J. P. Moreira, S. J. Ferreira, and R. H. C. Takahashi. 2007. Dual graph spatial cluster detection for syndromic surveillance in networks**.** *Advances in Disease Surveillance* 4: 88-92.

Duczmal, L., A. L. F. Cancado, Ricardo H.C. Takahashi, and Lupercio F. Bessegato. 2007. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis* 52, no. 1: 43-52.

Duda, R. O., D. G. Stork, and P. E. Hart. 2000. *Pattern Classification*. New York: Wiley.

Elgammal, A., R. Duraiswami, D. Harwood, and L. S. Davis. 2002. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceeding of IEEE* 90: 1151-1163.

ESRI. 2008. Download Census 2000 Tiger/line data. Available at: http://arcdata.esri.com/data/tiger2000/tiger_download.cfm. Accessed on January 2008.

Fischer, E. A. J., D. Pahan, S. K. Chowdhury, L. Oskam, and J. H. Richardus. 2008. The spatial distribution of leprosy in four villages in Bangladesh: an observational study. *BMC Infectious Disease* 8: 125-131.

Fraser, D. W., T. R. Tsai, W. Orenstein, W. E. Parkin, H. J. Beecham, R. G. Sharrar, J. Harris, G. F. Mallison, S. M. Martin, J. E. McDade, C. C. Shepard, and P. S. Brachman. 1977. Legionnaires' disease: Description of an epidemic of pneumonia. *The New England Journal of Medicine* 297(22): 1189-1197.

Fotheringham, A. S. and F. B. Zhan. 1996. A comparison of three exploratory methods for cluster detection in spatial point patterns. *Geographical Analysis* 28 (3): 200-218.

Gaines, K.F., A.L. Jr Bryan, and P. M. Dixon. 2000. The effects of drought on foraging habitat selection in breeding wood storks in coastal Georgia. *Waterbirds* 23: 64–73.

Getis, A. and J. K. Ord. 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24**:** 189-206.

Greenberg, M. and D. Wartenberg. 1991. Communicating to an alarmed community about cancer clusters: A fifty state survey. *Journal of Community Health* 16(2): 71-82.

Grimson, R. C. 1989. Assessing patterns of epidemiologic events in space-time. In Proceedings of the 1989 Public Health Conference on Records and Statistics. Hyattsville, MD: National Center for Health Statistics.

Haggett, Peter. 1965. *Locational analysis in human geography*. London: Edward Arnold.

Haining, R. 1998. Spatial statistics and the analysis of health data. In *Gis and health: Gisdata.* Anthony C. Gatrell and Markku Löytönen (Eds.). pp. 29-47. London: Taylor & Francis.

Heath, C. W. J. 1996. Investigating causation in cancer clusters. *Radiaion and Environmental Biophysics* 35(3): 133-136.

Hightower, A. W., M. Ombock, R. Otieno, and R. Odhiambo. 1998. A geographic information system applied to a malaria field study in western Kenya. *American Journal of Tropical Medicine and Hygiene* 58: 266-272.

Hjalmars, U., M. Kulldorff, G. Gustafsson, and N. Nagarwalla. 1996. Childhood leukaemia in Sweden: using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine* 15: 707-715.

Holden, Edgar. *Mortality and Sanitary Record of Newark, N.J. from 1859 to 1879*. A report presented to the president and directors of the Mutual Benefit Life Insurance Co., January, 1880.

Holmberg, P. C., S. Hernberg, K. Kurppa, K. Rantala, and R. Riala. 1982. Oral clefts and organic solvent exposure during pregnancy. *International Archives of Occupational and Environmental Health* 50(4): 371-376.

Inselberg, A. 1985. The Plane with Parallel Coordinates. *The Visual Computer* 1: 69-91.

Jacquez, G. M. 2008. Spatial Cluster Analysis. In *The Handbook of Geographic Information Science*, S. Fotheringham and J. Wilson (Eds.). pp. 395-416. England: Blackwell Publishing.

Jones, A. P., I. H. Langford, and G. Bentham. 1996. The application of K-function analysis to the geographical distribution of road traffic accident outcomes in Norfolk, England. *Social Science and Medicine* 42: 879-885.

Klemz, Bruce R. and Patrick M. Dunne. 2000. Exploratory analysis using parallel coordinate systems: Data visualization in n-dimensions. *Marketing Letters* 11: 323-333.

Knox, E. G. 1989. Detection of Clusters. In *Methodology of Enquiries into Disease Clustering*. E. G. Knox (Eds.). pp. 45-76. London: Small Area Health Statistics Unit.

Kulldorff, M. and N. Nagarwalla. 1995. Spatial disease clusters: Detection and inference. *Statistics in Medicine* 14 : 799-810.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26 (6): 1481-1496.

Kulldorff, M. 1998. Statistical methods for spatial epidemiology: Tests for randomness. . In *GIs and health*. Anthony Gattrell and M Loytonen (Eds.). pp. 49-62. London: CRC Press LLC.

Kulldorff, M. 1999. An isotonic spatial scan statistic for geographical disease surveillance. *Bulletin of National Institute of Public Health* 48 (2): 94-101.

Kulldorff, M. 1999. Spatial scan statistics: Models, calculations, and applications. In *Scan statistics and applications*, J. Glaz and N. Balakrishnan (Eds.). pp: 303-322. Boston, USA: Birkhäuser.

Kulldorff, M. 2001. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series a-Statistics in Society* 164: 61-72.

Kulldorff, M., T. Tango, and P. J. Park. 2003. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* 42(4): 665-684.

Kulldorff, M. 2006. Tests of spatial randomness adjusted for an inhomogeneity: A general framework. *Journal of the American Statistical Association* 101(475): 1289-1305.

Kulldorff, M., L. Huang, L. Pickle, and L. Duczmal. 2006. An elliptic spatial scan statistic. *Statistics in Medicine* 25(22): 3929-3943.

Lancet. 1990. Disease clustering: Hide or seek? *Lancet* 336(8717): 717-718.

Lawson, A. B. 1993. On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society* 156: 363–377.

Lawson, A. B., and F. Williams. 1994. Armadale—a case-study in environmental epidemiology. *Journal of the Royal Statistical Society* 157: 285–298.

Lawson, A. B., and F. Williams. 2003. *An introductory guide to disease mapping*. England: John Wiley & Sons.

Laumon, B., J. L. Martin, I. Bertucat, M. P. Verney, and E. Robert. 1996. Exposure to organic solvents during pregnancy and oral clefts: a case-control study. *Reproductive Toxicology* 10(1): 15-19.

Leung M-Y, K. Choi, A. Xai, and L. Chen. 2005. Non-random clusters of palindromes in Herpesvirus genomes. *Journal of Computational Biology* 12, 331-354.

Marshall, R. J. 1991. A review of methods for the statistical-analysis of spatial patterns of disease. *Journal of the Royal Statistical Society Series a-Statistics in Society* 154: 421-441.

Minamisava R., S. S. Nouer, N. O. L. Morais, L. K. Melo, and A. L. S. Andrade. 2009. Spatial clusters of violent deaths in a newly urbanized region of Brail: highlight the social disparities. *International Journal of Health Geograhpics* 8: 66-76.

Moore, D. A., and T. E. Carpenter. 1999. Spatial analytical methods and geographic information systems: Use in health research and epidemiology. *Epidemiologic Reviews* 21(2): 143-161.

Moran, P. A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society (Methodological)* 10: 243-251.

Nakaya T., and K. Yano. 2010. Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* 14: 223-239.

Naus, J. I. 1965. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association* 60: 532-538.

Neill D. B., A. Moore, and M. Sabhanani. 2005. Detecting elongated disease cluster. *Morbidity and Mortality Weekly Report* 54: 197-205.

Neill, D. B. 2006. Detection of spatial and spatio-temporal clusters. Phd Dissertation. Carnegie Mellon University.

Neill, Daniel B. and Andrew w. Moore. 2004. A fast multi-resolution method for detection of significant spatial disease clusters. *Advances in Neural Information Processing Systems* 16: 651-658.

Neutra, R. R. 1990. Counterpoint from a cluster buster. *American Journal of Epidemiology* 132(1): 1-8.

Nuckols J. R., M. H. Ward, and L. Jarup. 2004. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* 1121: 1007-1015.

Nunna, N., K. Wu, I. M. Young, J. W. Crawford, and K. Ritz. 2002. In Situ Spatial Patterns of Soil Bacterial Populations, Mapped at Multiple Scales, in an Arable Soil. *Microbial Ecology* 44: 296-305.

Openshaw, S., M E. Charlton, C. Wymer, and A. Craft. 1987. Mark i geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1(4): 335-358.

Openshaw, S., A. W. Craft, M. Charlton, and J. M. Birch. 1988. Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet* 1(8580): 272-273.

Ord, J. K. and A. Getis. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27: 286-306.

Ozkaynak H., T. Palma, J. S. Touma, and J. Thurman. 2008. Modeling population exposures to outdoor sources of hazardous air pollutants. *Journal of Exposure Science and Environmental Epidemiology* 18: 45-58.

Patil, G. P. and C. Taillie. 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11(2): 183-197.

Peter, H. 1995. Spatial pattern analysis in ecology based on Ripley's K-function: introduction and methods of edge correction. *Journal of Vegetation Science* 6: 575-582.

Pickle, L. W. 2002. Spatial analysis of disease. In *Biostatistical applications in cancer research*, C. Beam (Eds.). pp. 113-150. Boston: Klewer Academic Publisher.

Popovich, M. L. and B. Tatham. 1997. Use of immunization data and automated mapping techniques to target public health outreach programs. *American Journal of Preventive Medicine* 13: 102-107.

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

Patil, G. P. and C. Taillie. 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11, no. 2: 183-197.

Puett, R. C., A. B. Lawson, A. B. Clark, T. E. Aldrich, D. E. Porter, C. E. Feigley, and J. R. Hebert. 2005. Scale and shape issues in focused cluster power for count data. *International Journal of Health Geographics* 4: 1-16.

Puett, R. C., A. B. Lawson, A. B. Clark, J. R. Hebert, and M. Kulldorff. 2009. Power evaluation of focused cluster tests. *Environmental Ecological Statistics* 18: 1-14.

Robinson, T. P. 2000. Spatial statistics and geographical information systems in epidemiology and public health. *Advance in Parasitology* 47: 81-128.

Rogerson, P. A. 1997. Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine* 16: 2081-2093.

Rogerson, P. and I. Yamada. 2009. *Statistical detection and surveillance of geographic clusters*. New York: Taylor & Francis Group.

Rothman, K. J. 1990. A sobering start of the cluster busters conference. *American Journal of Epidemiology* 132: s6-s13.

Rushton, G. and P. Lolonis. 1996. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine* 15 (7-9): 717-26.

Sahajpal R., G. V. Ramaraju, and V. Bhatt. 2005. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *Conference on Knowledge Discovery in Data Mining* 1: 35-40.

Sankoh, O. A. and H. Becher. 2002. Disease cluster methods in epidemiology and application to data on childhood mortality in rural Burkina Faso.

Scott, D. W. 1992. *Mulivariate Density Estimation*. New York: Wiley- Interscience.

Skellam, J. G. 1952. Studies in statistical ecology: Spatial pattern. *Biometrika* 39, no. 3-4: 346-362.

Snow, John. 1855. *On the mode of communication of cholera.* London, UK: John Churchill.

Sun, Q. 2008. Statistical modeling and inference for multiple temporal or spatial cluster detection. Phd Dissertation. The State University of New Jersey.

Tango, T. 1995. A class of tests for detecting "general" and "focused" clustering of rare diseases. *Statistics in Medicine* 14: 2323-2334.

Tango, T. 2000. A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* 19: 191-204.

Tango, T. and K. Takahashi. 2005. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 4: 11-26.

Tonini M., D. Tuia, and F. Ratle. 2009. Detection of clusters using space-time scan statistics. *International Journal of Wildland Fires* 18: 830-836.

Trumbo, C. W. 2000. Public requests for cancer cluster investigations: A survey of state health departments. *American Journal of Public Health* 90(8): 1300-1302.

Turnbull, B. W., E. J. Wano, W. S. Burnett, H. L. Howe, and L. C. Clark. 1990. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 132: 136-143.

U. S. Census Bureau. 2000. Your gateway to census 2000. Available at: http://en.wikipedia.org/wiki/urbanization. Accessed on August, 2008.

Wakefield, J., M. Quinn, and G. Raab. 2001. Editorial: Disease clusters and ecological studies. *Journal of the Royal Statistical Society Series a-Statistics in Society* 164: 1-2.

Waller, L. A., B. W. Turnbull, L. C. Clark, and P. Nasca. 1992. Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE31 Spatial Cluster Analysis contaminated dumpsites in upstate New York. *Environmetrics* 3: 281-300.

Walter, S. D. 1992. The analysis of spatial patterns of health data I: the power to detect environmental effects. *American Journal of Epidemiology* 136: 742–759.

Wartenberg, D. 2001. Investigating disease clusters: Why, when and how? *Journal of the Royal Statistical Society Series a-Statistics in Society* 164: 13-22.

Wartenberg, D. and M. Greenberg. 1993. Solving the cluster puzzle - clues to follow and pitfalls to avoid. *Statistics in Medicine* 12(19-20): 1763-1770.

Wartenberg, D. 1995. Should we boost or bust cluster investigations. *Epidemiology* 6(6): 575-576.

Wen S., and B. Kedem. 2009. A semiparametric cluster detection method – a comprehensive power comparison with Kulldorff's method. *International Journal of Health Geographics* 8: 73-89.

Wieland, S. C., J. S. Brownstein, B. Berger, and K. D. Mandl. 2007. Density-equalizing euclidean minimum spanning trees for the detection of all disease cluster shapes. *Proc Natl Acad Sci U S A* 104, no. 22: 9404-9.

Wong, D. S. and J. Lee. 2005. *Statistical analysis of geographic information with arcview gis and arcgis*. Hoboken, N.J.: John Wiley & Sons, Inc.

Yiannakoulias, N., R. J. Rosychuk, and J. Hodgson. 2007. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics* 6: 28.

VITA

Zhijun Yao was born in Wuxi City, Jiangsu province in China on November 15, 1975, the son of Yuliang Cao and Huiying Yao. After graduated from Tianyi high school in 1995, he entered Nanjing University to study GIS and received the degree in Bachelor of Science in 1999. In that same year, he was accepted to the Master's degree program at Nanjing University. After graduated from Nanjing University with a Master of Science degree in GIS, he worked as GIS engineer in Jiangsu Telecom Science and Technology R&D Institute Co., Ltd in Nanjing for two years till he was enrolled to the PhD program in Geographic Information Science at Texas State in 2004. His dissertation research focused on the development of spatial clustering methods for health data analysis, particular in the detection of general shape disease clusters and cluster analysis involving multiple variables.

Permanent Address: #28 Henan Old Street, Zhangjing Town, XiShan District
Wuxi, Jiangsu Province, P. R. China 214194

This dissertation was typed by Zhijun Yao.