ABSTRACT

Title of dissertation: Semiparametric Cluster Detection

Shihua Wen
Doctor of Philosophy, 2007

Dissertation directed by: Professor Benjamin Kedem
Mathematical Statistics Program
Department of Mathematics

In this dissertation, a Semiparametric density ratio testing method which borrows strength from two or more samples is applied to moving windows of variable size in cluster detection. This Semiparametric cluster detection method requires neither the prior knowledge of the underlying distribution nor the number of cases before scanning. To take into account the multiple testing problem induced by numerous overlapping windows, Storey's q-value method, a false discovery rate (FDR) methodology, is used in conjunction with the Semiparametric testing procedure.

Monte Carlo power studies show that for binary data, the Semiparametric cluster detection method and its competitor, Kulldorff's scan statistics method, both achieve similar high power in detecting unknown hot-spot clusters. When the data are not binary, the Semiparametric methodology is still applicable, but Kulldorff's method may not be as it requires the choice of a correct probability model, namely the correct scan statistic, in order to achieve power comparable to that achieved by the Semiparametric method. Kulldorff's method with an inappropriate probability model may lose power.

Moreover, when the data are binary, the Semiparametric density ratio model reduces to the same scan statistic as Kulldorff's Bernoulli model. If a cluster candidate is known, under certain conditions the Semiparametric method achieves a higher power than the power achieved by a certain focused test in testing the hypothesis of no cluster.

The Semiparametric method potential in cluster detection is illustrated using a North Humberside childhood leukemia data set and a Maryland-DC-Virginia crime data set.

Semiparametric Cluster Detection

by

Shihua Wen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Larry Davis
Professor Laura Dugan
Professor Galit Shmueli
Professor Paul Smith

Dedication

For my parents.

Acknowledgments

With sincere gratitude, I want to thank all the people who helped and encouraged me through my graduate study. I also want to thank God or Buddha, they always come at the last minute but still in time to give me inspiration for solving the difficulties I met.

Special thanks to my adviser, Professor Benjamin Kedem. This research could not be done without his correct guidance, sincere encouragement, valuable hints, and tremendous help in my writing. I appreciate him for giving me such an interesting topic and for introducing me to a wonderful Semiparametric idea. I consider myself lucky working with him. I enjoy all the talking with him, and his humor impresses me a lot.

I also want to thank all my committee members. Professor Paul Smith is always nice to everybody. I learned Linear Models from him, and thank him for giving me the opportunity to work in the STAT lab. It was a great experience. Professor Galit Shmueli gave me great help and encouragement in this research. I thank her for introducing me to other current methods in scan statistics and cluster detection. I also appreciate her sharing research papers and other research information with me. Professor Laura Dugan is an expert in criminology and gave me a lot of valuable information on crime research as well as useful data sets. I am grateful for her generous help and her willingness to serve on my committee. Professor Larry Davis is a well known scholar in computer science. It is my great honor to have Professor Davis as the Dean's Representative in my dissertation committee.

I would like to extend my thanks to Dr. Tiwari for serving on my candidacy committee, his comments on my research, and for arranging a wonderful opportunity to talk at National Institute of Cancer (NCI/NIH). I also wish to express my gratitude to Professor Martin Kulldorff at Harvard University and to Professor Reza Modarres at George Washington University for their important comments and suggestions.

Finally, I want to thank my parents for their endless support and trust, my friends for their great help and encouragement. Thanks a lot for everybody again. This study would never have been accomplished without all of these people. I really appreciate it.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this dissertation, I develop a semiparametric scan statistics method for cluster detection. I refer to this method as *semiparametric cluster detection method* or *semiparametric method* in short. This method applies a semiparametric density ratio model to moving windows of variable size to scan the study region and detect potential clusters of events or cases. The simulation studies show that the statistical power of the semiparametric method is comparable to the current Kulldorff's spatial scan statistics method [29, 30, 31], but the semiparametric method requires fewer distributional assumptions on the data. The semiparametric method works well in many cases adhering to a unified setting [26, 70], but Kulldorff's method requires the choices of a correct probability model, namely the correct scan statistic, in order to achieve the power achieved by the semiparametric method. The semiparametric scan statistics methodology has also been successfully applied to real data which points to its potential in cluster detection. The first chapter gives a brief description of the purpose and the general frame of this dissertation.

1.1 Scan Statistics and Cluster Detection

Scan statistics arise when scanning in time or space, or both, looking for unusual clusters of certain events or cases [18]. Here, an *event* can be the occurrence

of some type of disease, or some sort of physical or chemical measurements, etc. A *cluster* is defined as a certain spatial or temporal subregion where the probability distribution of an event is different from the event probability distribution in the rest of the region. More generally, a cluster is a subregion where the behavior of an observable is different from the behavior of the observable in the rest of the region. For instance, a city neighborhood where the crime rate is higher than in the rest of the city defines a cluster. Another example can be a subregion comprised of several counties with higher disease rate than all other counties in a region. If we can locate (detect) the clusters more accurately, we can make better decisions and more efficient policies.

The modern literature about scan statistics can be traced back to the 1960's. Since then, it has been applied in many fields, including epidemiology, criminology, economics, health management, brain imaging, genetics, mining, quality control, astronomy, syndromic surveillance, and so on. See Glaz et al. (2001), Glaz and Naus (1991), Kulldorff (1999), Naus (1965), Pickle et al. (2003), and Shmueli et al. (2006) [18, 43, 30, 49, 58]. Of particular importance is the so called Kulldorff's spatial scan statistics method. Kulldorff's method uses circular, elliptic, or cylindrical scan window to detect clusters in two or higher dimensions, their location and size, by making an assumption about the underlying distribution (typically Bernoulli or Poisson) of the scanned region. These distributional assumptions are used in computing Kulldorff's spatial scan statistic, a likelihood ratio type test statistic, to determine the cluster candidate. A p-value for the significance of the cluster candi-

date is obtained by a Monte Carlo hypothesis testing procedure or by a permutation test [29, 30, 31]. A detailed review of Kulldorff's method is given in Chapter 2.

In this dissertation we propose a certain semiparametric generalization of Kulldorff's method which requires much less than complete distributional assumptions and which does not require the number of cases prior to scanning for the time consuming Monte Carlo hypothesis testing. The semiparametric approach used in this research is the density ratio model as in Fokianos et al. (2001), Qin and Lawless (1994), and Qin and Zhang (1997) [13, 52, 53]. Given $m = q + 1$ samples,

$$\frac{g_j(x)}{g_m(x)} = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)\}, \qquad j = 1, \ldots, q, \quad q = m - 1$$

where $g_m(x) \equiv g(x)$ is the (reference) probability density function of the mth sample, $g_j(x)$ is the probability density function of the jth sample, $(\alpha_j, \boldsymbol{\beta}_j')$ are the parameters relating the jth and the reference densities, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})'$ with dimension p depends on the choice of the known *tilt* vector-valued function $\boldsymbol{h}(x)$. By following this setup, testing for distributional homogeneity is equivalent to testing H_0: $\boldsymbol{\beta}$=$\mathbf{0}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$. Other than an assumption concerning the tilt function $\boldsymbol{h}(x)$, this method does not require prior knowledge of any distribution. Once $\boldsymbol{h}(x)$ is chosen, all the parameters and the reference distribution function $G(x)$ are estimated from the combined data composed of all the samples.

The semiparametric cluster detection method discussed in this dissertation merges the semiparametric density ratio model with Kulldorff's scan procedure leading to a fairly general cluster detection procedure. The idea is quite natural. Since cluster detection amounts essentially to testing the homogeneity of the probability

distributions between the cluster region and the non-cluster region, the cluster detection problem is to test if $\boldsymbol{\beta} = 0$ in the semiparametric density ratio model. A detailed study of this semiparametric method is described in Chapter 3.

1.2 Other Scan Statistics Methods

Besides Kulldorff's method and the semiparametric method proposed in this dissertation, there are other types of scan statistics methods. I list some of them in this section.

Glaz et al. (2001) defines a scan statistic S_w based on point data where the occurrence of an event is represented by a point in the study interval. This interval could be a one dimensional line, such as time, or a higher dimensional set, such as a geographic map [18]. The point data are assumed to be distributed uniformly or to follow a Poisson process over the whole study interval. Figure 1.1 gives an example of this type of scan statistic. The solid line represents a time line scaled into [0, 1). Each small triangle (point) is denoted as an event occurring at that moment. A scan window slides along the line. Let S_w be the largest number of events occurred in a window of fixed size w. Then this S_w is called the scan statistic and the corresponding window is the cluster candidate. The problem of interest is the probability of k or more events occurred in the given window w. More precisely, if the total number of N events over the interval is given, the problem is to compute the retrospective probability $P(S_w \geq k|N)$, which is a conditional probability. If N is viewed as a random variable, on the other hand, the prospective probability

$P(S_w \geq k)$ is unconditional.

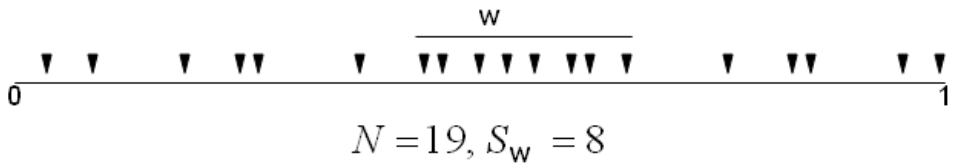


Figure 1.1: Illustration of Glaz's scan statistic in one dimension.

Given N points independently uniformly distributed on interval [0,1), Wallenstein and Neff (1987) [67] gives the following easy to compute approximation for $P(S_w \geq k|N)$ as a simple sum of binomial and cumulative binomial probabilities. Let

$$b(k; N, w) = \left(\begin{array}{c} N \\ \\ k \end{array} \right) w^k (1-w)^{N-k}$$

$$G_b(k; N, w) = \sum_{i=k}^{N} b(i; N, w)$$

Then approximately,

$$P(S_w \geq k|N) \approx \left(\frac{k}{w} - N - 1 \right) \cdot b(k; N, w) + 2G_b(k; N, w) \tag{1.1}$$

When the data follows a Poisson Process on the interval [0,T) with rate λ, Newell (1963) [46] gives the following asymptotic formula of $P(S_w \geq k)$,

$$P(S_w \geq k) \approx 1 - \exp\{-\lambda^k w^k T/(k-1)!\}$$

Glaz and Naus (1991) give tight bounds and approximations for scan statistic probabilities for independently and identically distributed (i.i.d.) discrete data for

fixed window size [17]. Naus and Wallenstein (2004) derive accurate approximations for the joint distributions of scan statistics for a range of values of w, or of k, that can be used to set an experiment-wide level of significance that takes into account the multiple comparisons involved. This makes it possible to determine the cluster sizes from various scanning window sizes [45].

Pozdnyakov et al. (2004) propose a martingale method for binary data to approximate the distributions of a wide variety of scan statistics, including some for which analytical results are computationally infeasible [50]. Glaz and Zhang (2004) derive multiple scan statistics of variable window sizes for i.i.d. Bernoulli trials (0/1)in one or two dimensional intervals. They also derive simple approximations for the significance level of the scan statistics [19]. Glaz and Zhang (2006) propose a maximum scan score-type statistic for testing the null hypothesis that the observations are i.i.d. according to a specified distribution, against an alternative that the observations cluster within a window of unknown length. This statistic is a variable window scan statistic, based on a finite number of standardized fixed window scan statistics. Approximations for the significance level of this statistic are derived for $0-1$ i.i.d. Bernoulli trials uniformly distributed in the interval [0, 1).

Kulldorff's method and the semiparametric method we propose in this dissertation use circular or regular shape scanning windows. Patil and Taillie (2004) propose a upper level set (ULS) scan method [47, 48] which detects hot-spot clusters with irregular shape. The main idea of the ULS method is that the whole study region is composed of cells with rate (or intensity) $G_a = Y_a/A_a$, where Y_a is the raw

count and A_a is the "size" of cell a. A zone Z is a union of connected cells and Ω is a collection of all the possible Z's. For a given g, define an upper level set as $U_g = \{a : G_a \geq g\}$. A reduced space Ω_{ULS} is a collection of all the possible unions of the connected cells in U_g. Thus, all the zones, $Z \in \Omega_{ULS}$, can possibly become scan windows and can be of any shape. Once the scan window is determined, similar to Kulldorff's method, a probability assumption is imposed to the study region to get a likelihood ratio type statistic and the p-value of the cluster candidate is obtained by Monte Carlo simulation. Modarres and Patil (2006) extend this ULS methodology to bivariate data [42]. Tango and Takahashi (2005) propose a flexibly shaped scan method [65]. It imposes an irregularly shaped window Z on each cell (e.g. county) by connecting its adjacent cells and computes the likelihood ratio type statistic as in Kulldorff's method.

Most of the above mentioned methods apply to point data where the occurrence of an event is represented by a point and those points are assumed to be distributed uniformly or follow a Poisson process in the study interval. Kulldorff's method and the ULS method can handle non-binary data, but still need to assume a certain probability model. The semiparametric cluster detection method proposed in this dissertation, however, does not require those specific assumptions and is applicable to many data types. See Chapter 3 for the details of the semiparametric method.

The above scan statistics methods are mainly used to detect the location, the size and the significance of local clusters. If the hypothesis is that the risk in a

specified region is higher than in the rest of the region, for example, the risk of a type of disease is higher close to a nuclear power plant than in the rest of the area, then *focused cluster tests* are used. I will briefly describe the Lawson-Waller focused test [68] in Chapter 4 when we conduct the power study. Sometimes researchers are interested in evaluating the presence of clustering throughout the study region. For example, we might want to know if a particular disease is infectious or not, in which case we would expect cases to be found close to each other no matter where they occur. In this case, global clustering tests should be used, such as Cuzick-Edwards' (1990) k nearest neighbor (k-NN) method [6], Tango's (2000) maximized excess events test (MEET), Bonetti-Pagano's (2005) M-statistic [4], and so on. Since the semiparametric method is a cluster detection method aiming to detect the local clusters, these global clustering tests are out of the focus of this dissertation.

In surveillance or quality control fields, early detection of outbreaks is essential for successful operation and prevention of disasters. In the purely temporal setting, traditional control charts method, including Shewhart charts, moving average charts, Cumulative sum (CumSum) charts, etc., and time series methods are used. Recently wavelet-based methods were found to offer a more elegant and suitable solution for early detection. If multiple data sets or streams are present, the multivariate versions of the control charts, time series, and wavelet-based methods could be potentially implemented. Shmueli and Fienberg (2006) gives a nice review about the these options [58].

Naus and Wartenberg (1997) have developed purely temporal scan statistics for

two data types with the purpose of finding clusters with a minimum number of both types of events [44]. In addition, Kulldorff et al. (2007) develop a multivariate scan model which simultaneously incorporates multiple data sets into a single likelihood function to search for clusters. Chapter 2 gives a brief description of Kulldorff's multivariate scan model.

1.3 Dissertation Map

This dissertation is organized as follows. Chapter 2 describes Kulldorff's scan statistics method, including the Bernoulli model, Poisson model, ordinal model, exponential model, and so on. Chapter 3 introduces the semiparametric density ratio model and our semiparametric cluster detection method. Chapter 4 presents power studies comparing Kulldorff's method and our semiparametric method, including a limited power study given that the location and the size of a cluster candidate are known, and a complete power study which where the cluster candidate is unknown. Chapter 5 illustrates the cluster detection potential of the semiparametric method by analyzing a North Humberside childhood leukemia data set [26, 1] and a Maryland-DC-Virginia crime data set. Chapter 6 summarizes the whole dissertation and discusses possible improvements of the semiparametric cluster detection method for future research.

Chapter 2

Kulldorff's Scan Statistics Method

2.1 Overview

A number of different tests for detecting spatial clusters, temporal and spatial, have been proposed in the last three decades. One of the most popular methods is Kulldorff's scan statistics. Kulldorff's scan statistics method can detect both the location and the size of a cluster simultaneously by using a large collection of overlapping scan windows [29, 30]. For spatial data, the method first imposes a circular scan window on a map and lets the circle centroid move across the study region. For any given centroid, the radius of the window varies continuously from zero to some upper limit. Usually this upper limit is set to be the radius which covers 50% of the whole study region or population. In this way, the method generates a large set of scan windows $\mathbb{Z}$ with different centroids and sizes. Under the null hypothesis of no cluster, the underlying behavior of the data throughout the whole study region is the same. Under the alternative hypothesis, there is at least one scan window for which the underlying behavior is different inside the window as compared with its complement, which means any scan window Z could be a potential cluster. In practice, some data are updated periodically, Kulldorff (2001) [31] suggested space-time scans for such cases. The scanning procedure of space-time scans is almost identical to the purely spatial scan, except that the scan window becomes

a three dimensional cylinder instead of two dimensional. See Figure 2.1. Since the statistical formulation of space-time scan is identical to the two dimensional case, we will only discuss the two dimensional purely spatial scan in this paper.

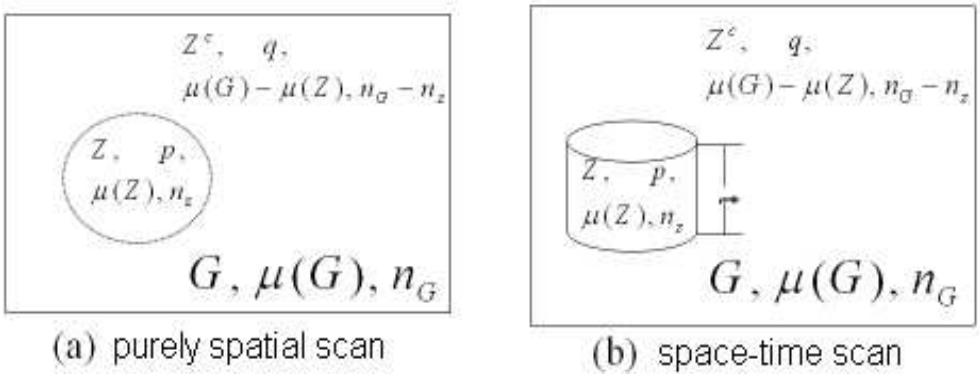


Figure 2.1: Notation for Kulldorff's method

In Kulldorff's scan statistics method, each scan window Z is associated with a likelihood ratio test statistic $\lambda(Z)$ which can be computed based on the chosen underlying probability model and the observed data inside and outside the scanning window. The scan window associated with the maximum $\lambda(Z)$ is defined as the primary cluster candidate occurring not by random chance. The maximum likelihood ratio itself is called the Kulldorff's spatial scan statistic, and the null hypothesis is rejected for large value of the statistic. After the spatial scan statistic and the primary cluster candidate are determined, a Monte Carlo hypothesis procedure [10] or a permutation test procedure [33, 22] is executed to generate the probability distribution of Kulldorff's scan statistic under the null hypothesis of no cluster in the study region, and a p-value is obtained. The detail steps of the Monte Carlo hypothesis testing procedure is as follows.

1. Obtain the value of Kulldorff's scan statistic for the true data at hand.

2. Given the total number of the cases (events), create a large number of random data sets generated under H_0 for the whole study region.

3. Calculate the value of Kulldorff's scan statistic for each random replication.

4. Sort the values of the Kulldorff's scan statistics from the true and the generated data sets, and note the rank of the one calculated from the true data set to obtain the p-value.

The following sections describe Kulldorff's scan statistics for the Bernoulli model, Poisson model, and ordinal model [29, 30, 23]. As for other types of scan statistics in Kulldorff's scan statistics family, see Huang et al. (2007) for the exponential model [22], Kulldorff et al. (2007) for the multivariate scan model [36], Kulldorff et al. (2006a) for the normal model [35], and Kulldorff et al. (2006b) for elliptic window scans [34].

2.2 Bernoulli Model

Bernoulli-based scan statistics are used when individual entities have only two states such as an individual person having breast cancer or not. Figure 2.1(a) shows a typical setup of Kulldorff's scan statistics method. The "purely spatial scan" means a two-dimensional scan on a geographical map.

- G: the whole study region.

- Z : the scan window.
- Z^C : outside scan window.
- $\mu(G)$: the total number of individual. entities (e.g. people) in G.
- $\mu(Z)$: the number of individual entities in Z.
- n_G : the total number of events in G.
- n_Z : the number of events inside Z.
- p : the rate of events that occurred inside the scan window Z.
- q : the rate of events that occurred outside the scan window Z.

Clearly, $\mu(Z^c) = \mu(G) - \mu(Z)$ and $n_{Z^C} = n_G - n_Z$. We test the null hypothesis $H_0 : p = q$ of no cluster. The following shows that for an alternative hypothesis that there is a hot spot cluster $p > q$, each scan window Z invokes a likelihood ratio test statistic as in equation (2.1).

Consider binary $0-1$ data. The likelihood for a fixed scan window Z is

$$L(Z, p, q) = p^{n_Z}(1-p)^{\mu(Z)-n_Z} \times q^{n_G-n_Z}(1-q)^{(\mu(G)-\mu(Z))-(n_G-n_Z)}.$$

Under $H_0 : p = q$, $\hat{p}_0 = n_G/\mu(G)$, and the maximized likelihood becomes L_0 which is independent of Z,

$$L_0 = \sup_{H_0:p=q} L(Z, p, q) = \hat{p_0}^{n_G}(1-\hat{p_0})^{\mu(G)-n_G}.$$

Under $H_A : p > q$, $\hat{p} = n_Z/\mu(Z)$, $\hat{q} = (n_G - n_Z)/\left(\mu(G) - \mu(Z)\right)$, and the likelihood $L(Z)$ is a function of Z,

$$\begin{aligned} L(Z) &= \sup_{H_A: p>q} L(Z, p, q) \\ &= \hat{p}^{n_Z}(1-\hat{p})^{\mu(Z)-n_Z} \times \hat{q}^{n_G-n_Z}(1-\hat{q})^{(\mu(G)-\mu(Z))-(n_G-n_Z)}. \end{aligned}$$

Therefore, the likelihood ratio for a fixed scan window Z is

$$\begin{aligned} \lambda(Z) &= \frac{\sup_{H_A: p>q} L(Z, p, q)}{\sup_{H_0: p=q} L(Z, p, q)} \\ &= \begin{cases} \frac{\hat{p}^{n_Z}(1-\hat{p})^{\mu(Z)-n_Z} \times \hat{q}^{n_G-n_Z}(1-\hat{q})^{(\mu(G)-\mu(Z))-(n_G-n_Z)}}{\hat{p}_0^{n_G}(1-\hat{p}_0)^{\mu(G)-n_G}} & \text{if } \hat{p} > \hat{q} \\ 1 & \text{otherwise} \end{cases}. \end{aligned} \tag{2.1}$$

If we were scanning for cold spots, then "$>$" would change to "$<$" above; if we were scanning for either hot or cold spots, then it would be "$\neq$" [32]. After all the $\lambda(Z)$ are computed, we determine the maximum of $\lambda(Z)$'s

$$\lambda = \sup_{Z \in G} \lambda(Z) \equiv \lambda(\hat{Z})$$

as Kulldorff's scan statistic and $\hat{Z}$ to be the primary cluster candidate. We reject the null hypothesis for large values of λ, and a Monte Carlo based p-value can be obtained from randomization of the cases across the whole study region, given the total number of cases as described in the previous section.

2.3 Poisson Model

Poisson-based scan statistics are used for the comparison of the number of cases inside and outside a scan window when searching for clusters. The notion and

setup of Kulldorff's scan under the Poisson model follow the same scheme shown in Figure 2.1(a). Suppose a study region is composed of I sub-regions. Assume the number of events x_i which occur in an "interval" $\mu(A_i)$ is a Poisson process with intensity rate p inside the scan window and q outside, where $i = 1, 2, \ldots, I$. For example, x_i could be the number of people with certain type of cancer in county A_i whose population size is $\mu(A_i)$. In addition, we have

$$n_G = \sum_{A_i, x_i \in G} x_i, \quad \mu(G) = \sum_{A_i \in G} \mu(A_i)$$

and

$$n_Z = \sum_{A_i, x_i \in Z} x_i, \quad \mu(Z) = \sum_{A_i \in Z} \mu(A_i).$$

The null hypothesis is still $H_0 : p = q$ and the method parallels the Bernoulli case.

For a given fixed scan window Z, the likelihood is

$$L(Z, p, q) = \frac{e^{-p \cdot \mu(Z)} (p \cdot \mu(Z))^{n_Z}}{n_Z!} \times \frac{e^{-q \cdot (\mu(G) - \mu(Z))} (q \cdot (\mu(G) - \mu(Z)))^{n_G - n_Z}}{(n_G - n_Z)!}.$$

Under $H_0 : p = q$, $\hat{p}_0 = n_G / \mu(G)$, we obtain

$$L_0 = \sup_{H_0 : p = q} L(Z, p, q) = \frac{e^{-n_G} \cdot (\frac{n_G}{\mu(G)})^{n_G} \cdot \mu(Z)^{n_Z} \cdot (\mu(G) - \mu(Z))^{n_G - n_Z}}{n_Z! \cdot (n_G - n_Z)!}.$$

Under $H_A : p > q$, $\hat{p} = n_Z / \mu(Z) \quad \hat{q} = (n_G - n_Z) / (\mu(G) - \mu(Z))$, we obtain

$$L(Z) = \sup_{H_A : p > q} L(Z, p, q) = \frac{e^{-n_G} \cdot (n_Z)^{n_Z} \cdot (n_G - n_Z)^{n_G - n_Z}}{n_Z! \cdot (n_G - n_Z)!}.$$

Therefore, the likelihood ratio for the scan window Z is shown in equation (2.2):

$$\lambda(Z) = \begin{cases} (\frac{n_Z}{e_Z})^{n_Z} \cdot (\frac{n_G - n_Z}{n_G - e_Z})^{(n_G - n_Z)} & \text{if } n_Z > e_Z \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

where n_Z is the observed number of cases inside the scan window Z and e_Z is the expected number of cases inside the scan window Z under the null hypothesis of no cluster. As before, if we were scanning for a cluster other than hot spot, we simply change the inequality sign as needed. After all the $\lambda(Z)$ are obtained, put $\lambda = \max_Z \lambda(Z) \equiv \lambda(\hat{Z})$, where $\hat{Z}$ is the primary cluster candidate, and reject the null hypothesis for large λ. A Monte Carlo based p-value can be obtained from randomization of the cases across the study region given the total number of cases, as in the Bernoulli model.

2.4 Ordinal Model

An ordinal model is used when individual entities have $K \geq 2$ ordinal categories such as the different stages of prostate cancer (Klassen, 2005) [28]. A higher category may reflect a more serious cancer stage. With the ordinal model, each observation is a case, and each case belongs to one of several ordinal categories. Suppose the study region consists of I sub-regions and the variable of interest is recorded in K categories. Let c_{ik} be the number of individuals in location i who fall into category k, where $i = 1, 2, \ldots, I$, and $k = 1, 2, \ldots, K$. Let $C_k = \sum_i c_{ik}$ be the number of observations in category k across the study region, and $C = \sum_k C_k = \sum_k \sum_i c_{ik}$ be the total number of observations in the whole study region. The null hypothesis of no cluster in this model means $p_1 = q_1, \ldots, p_k = q_k$, where p_k and q_k are the unknown probabilities that an observation belongs to category k inside and outside the scanning window, respectively. To detect subregions with high rates of higher

stages as compared with the rest of the area, one possible alternative hypothesis could be

$$\frac{p_1}{q_1} \leq \frac{p_2}{q_2} \leq \frac{p_K}{q_K}$$

searching for hot-spot clusters with an excess of cases in the high-valued categories. Obviously when $K = 2$, the ordinal model set up reduces to the Bernoulli model. Following similar scan procedures as in the Bernoulli and Poisson models, the likelihood ratio test statistic for each scan window is:

$$\lambda(Z) = \begin{cases} \prod_{k=1}^{K} \left[\prod_{i \in Z} \hat{p_k}^{c_{ik}} \cdot \prod_{i \notin Z} \hat{q_k}^{c_{ik}} \right] \Big/ \prod_{k=1}^{K} \left(\frac{C_k}{C}\right)^{C_k} & \text{otherwise} \\ 1 & p_k = q_k,\ k = 1, 2, \ldots, K \end{cases} \tag{2.3}$$

where $\hat{p}_k$ and $\hat{q}_k$ are the MLEs of p_k and q_k under the alternative hypothesis. A "Pool-Adjacent-Violators" algorithm can be applied to compute $\hat{p}_k$ and $\hat{q}_k$ [2, 11]. It is also possible to search for cold-spot clusters with an excess of cases in the low-valued categories or simultaneously for both hot or cold spots by reversing the order of the categories. After all the $\lambda(Z)$ are obtained, compute $\max_Z \lambda(Z) \equiv \lambda(\hat{Z})$, where $\hat{Z}$ is the primary cluster candidate, and reject the null hypothesis for large $\lambda(\hat{Z})$. A Monte Carlo based p-value can be obtained by randomization of the observations across the study region given the total number of observations in each category $(C_1, C_2, \ldots, C_K)$. For more detail about the ordinal model, see Jung et al. (2007) [23].

2.5 Other Models

There are other types of scan statistics in Kulldorff's scan statistics family, such as the exponential model mainly for survival time data, the normal model for continuous data that takes both positive and negative values, the multivariate scan model for analyzing multiple surveillance data sets simultaneously, and so on.

Exponential model: The exponential model is mainly designed for survival time data, and the likelihood function for the scan statistic is based on the exponential distribution. However, it also could be used for other positive continuous type data as well, especially for data with a heavy right tail. In the exponential model, each observation is a case, and each case has one continuous variable attribute as well as a 0/1 censoring designation. For survival data, the continuous variable is the time between diagnosis and death or, depending on the application, between two other types of events. If some of the data are censored, due to loss of follow-up, then the continuous variable is the time between diagnosis and time of censoring. The 0/1 censoring variable is used to distinguish between censored and non-censored observations. For more details about the exponential model and its scan statistic, see Huang et al. (2007) [22].

Normal model: The normal model is designed for continuous data and the likelihood function for the scan statistic is based on the normal distribution. For each individual, called a case, there is a single continuous attribute that may be either negative or positive. For example, the data may consist of the birth weight and residential census tract for all newborns, with an interest in finding clusters with

lower birth weight. The model can also be used for ordinal data when there are very many categories. That is, ties are allowed. It is also noticed that the results from the normal model can be greatly influenced by extreme outliers, so it may be wise to truncate such observations before doing the analysis. For more detail about the normal model and its scan statistic, see Kulldorff et al. (2006a) [35].

Multivariate scan: Sometimes, especially in disease surveillance, the statistical power to detect an outbreak that is present in all data sets may suffer due to low numbers in each data set. Kulldorff's Multivariate scan model can simultaneously incorporate multiple data sets into a single likelihood function searching for clusters and hence increasing the power. This could be done by defining the combined log-likelihood as the sum of the individual log-likelihoods for those data sets for which the observed case count is more than the expected, if hot-spot clusters are of interest. When searching for clusters with low rates, the same procedure is performed, except that we instead sum up the log-likelihood ratios of the data sets with fewer than the expected number of cases within the window in question. When searching for both high and low clusters, both sums are calculated, and the maximum of the two is used to represent the log likelihood ratio for that window. In multivariate scan, all data sets must use the same probability model and the same geographical coordinates file. For more detail, see Kulldorff et al. (2007) [36].

Elliptic window scans: The above Kulldorff's scan statistics commonly use a circular scanning window. To have more flexibility, the elliptic version of the Kulldorff's scan statistics uses a scanning window of variable location, shape (eccentric-

ity), angle and size, and with and without an eccentricity penalty. The mathematical principles behind the scan are identical for circular, elliptic or any other shape of the window, with the only difference being the collection of candidate cluster areas considered. In general, the elliptic scan statistic performs well for circular clusters, and equally important, the circular scan statistic performs well for elliptic clusters also. One possible advantage of the elliptic versus the circular scan statistic is that the former may give a better estimate of the true cluster area especially when the true cluster is an elongated one. But the circular scan statistic requires fewer computing resources. For more detail, see Kulldorff et al. (2006) [34].

Chapter 3

Semiparametric Scan Statistics Method

3.1 Overview

We have described Kulldorff's scan statistics method in the previous Chapter. For different types of data, Kulldorff's method needs to choose different probabilistic models. In this dissertation, we propose a semiparametric scan statistics method [26]. It uses the same scanning scheme as Kulldorff's, but a semiparametric method to develop the scan statistics and test the significance of the cluster candidate. To take into account the multiple testing problem induced by numerous overlapping windows, Stoney's q-value method [60, 61], a false discovery rate (FDR) methodology [3], is used in conjunction with the semiparametric testing procedure [70].

In the first section of this chapter, I first introduce the semiparametric density ratio model used in this research. Secondly, I discuss how the semiparametric density ratio model is applied to cluster detection and its advantages. In the last section of this chapter, I describe the concept of FDR methodology as well as the q-value used in this work.

3.2 Semiparametric Density Ratio Model

3.2.1 The Model

The Semiparametric method we use here is based on a density ratio model studied by Fokianos et al. (2001), and Qin and Zhang (1997) [13, 53]. Consider m independent samples,

$$\boldsymbol{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})' \sim g_1(x)$$

$$\boldsymbol{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2})' \sim g_2(x)$$

$$\vdots$$

$$\boldsymbol{x}_m = (x_{m1}, x_{m2}, \ldots, x_{mn_m})' \sim g_m(x)$$

where $g_j(x)$ is the probability density function of $x_{ji}, j = 1, \ldots, m,\ i = 1, \ldots, n_j$. Choosing the mth sample as the reference sample and $g_m(x)$ as the reference density, it is assumed that the density ratio between the jth density and the reference density has an exponential form as in (3.1),

$$\frac{g_j(x)}{g_m(x)} = \exp\{(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{h}(x))\}, \qquad j = 1, \ldots, q, \quad q = m - 1 \tag{3.1}$$

Notice that $\boldsymbol{h}(x)$ is a known function of x which may take on a scalar form such as x, x^2, or $\log x$, or a vector-valued form such as $(x, x^2)'$, or $(x, \log x)'$, and so on. See more in subsection 3.2.2. Here $\alpha_j = \alpha_j(\boldsymbol{\beta}_j)$ is a scalar, but $\boldsymbol{\beta}_j$ could be a scalar or vector depending on $\boldsymbol{h}(x)$. Clearly, $\boldsymbol{\beta}_j = \mathbf{0}$ implies $\alpha_j = 0$, and the hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_q = \mathbf{0}$ implies all the m samples come from a common distribution with probability density $g_m(x) \equiv g(x)$. In this section we shall assume

that function $\boldsymbol{h}(x)$ is p-dimensional; notice that p and q here are different from p and q in the previous chapter.

An example of the exponential density ratio model (3.1) is provided by multinomial logistic regression upon an appeal to Bayes theorem. Consider a categorical random variable y such that $P(y = j) = \pi_j$, where $f(x|y = j) = g_j(x), \quad j = 1, \ldots, m$, and $\sum_{i=1}^{n} \pi_j = 1$. If

$$P(y = j|x) = \frac{\exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)\}}{1 + \sum_{k=1}^{q} \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(x)\}}, \quad j = 1, 2, ..., q, \quad q = m - 1$$

then by Bayes rule, model (3.1) holds with $\alpha_j = \alpha_j^* + \log(\pi_m/\pi_j), j = 1, 2, ..., q$

3.2.2 Choice of the Tilt Function

The semiparametric method requires choosing an appropriate tilt function $\boldsymbol{h}(x)$. A clue of how to choose a satisfactory $\boldsymbol{h}(x)$ for a given situation can be derived from common exponential families as we show in some examples with $m = 2$ below. More examples can be found in Kay and Little (1987) [24].

Bernoulli distribution: For Bernoulli(p), the density ratio is

$$\begin{aligned} \frac{g_1(x)}{g_2(x)} &= \frac{p_1^x(1-p_1)^{1-x}}{p_2^x(1-p_2)^{1-x}} \\ &= \exp\{\log\frac{1-p_1}{1-p_2} + (\log\frac{p_1}{p_2} - \log\frac{1-p_1}{1-p_2}) \cdot x\}. \end{aligned}$$

So we obtain

$$\alpha = \log\frac{1-p_1}{1-p_2}, \quad \beta = \log\frac{p_1}{p_2} - \log\frac{1-p_1}{1-p_2}, \quad h(x) = x,$$

and $p_1 > p_2 \Leftrightarrow \beta > 0$.

Poisson distribution: For Poisson(λ), similarly we have

$$\alpha = -(\lambda_1 - \lambda_2), \quad \beta = \log \frac{\lambda_1}{\lambda_2}, \quad h(x) = x$$

and $\lambda_1 > \lambda_2 \Leftrightarrow \beta > 0$.

Normal distribution: For N(μ, σ^2), unequal means and variances,

$$\begin{aligned} \alpha &= \ln(\frac{\sigma_2}{\sigma_1}) + \frac{\mu_2^2}{2\sigma_2^2} - \frac{\mu_1^2}{2\sigma_1^2}, \\ \boldsymbol{\beta} &= \begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} = \begin{pmatrix} (\mu_1\sigma_2^2 - \mu_2\sigma_1^2) \Big/ \sigma_1^2\sigma_2^2 \\ (\sigma_1^2 - \sigma_2^2) \Big/ 2\sigma_1^2\sigma_2^2 \end{pmatrix}, \\ \boldsymbol{h}(x) &= (x, x^2)' \end{aligned}$$

If $\sigma_1 = \sigma_2$, then $\mu_1 > \mu_2 \Leftrightarrow \beta > 0$, and $h(x) = x$.

3.2.3 Parameter Estimation of the Model

Let $\boldsymbol{t} = (t_1, t_2, ..., t_n)' = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_{n_m}')'$ denote the combined data from the m samples, and put $p_i = dG(t_i), \quad i = 1, 2, ..., n$ where $n = n_1 + \cdots + n_m$. Then the likelihood becomes

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp\{\alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{h}(x_{1j})\} \ldots \prod_{j=1}^{n_q} \exp\{\alpha_q + \boldsymbol{\beta}_q' \boldsymbol{h}(x_{qj})\}. \qquad (3.2)$$

Following a profiling procedure discussed in Fokianos et al. (2001), Qin and Lawless (1994), and Qin and Zhang (1997) [13, 52, 53], first express each p_i in terms of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and then substitute the p_i back into the likelihood to produce a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ only. When $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are fixed, the likelihood (3.2) is maximized by maximizing

only the product term $\prod_{i=1}^{n} p_i$, subject to the m constraints,

$$\sum_{i=1}^{n} p_i = 1,$$

$$\sum_{i=1}^{n} p_i[w_j(t_i) - 1] = 0,\ j = 1, \ldots, q$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_q)'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_q')'$, and $\omega_j(t) = \exp\{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{h}(t)\}$ [51].

The maximization employs the method of Lagrange multipliers, the first of which becomes $\lambda_0 = n$, and the rest are expressed by construction as $\lambda_j = \nu_j n$, $j = 1, \ldots, q$, for some ν_j. It follows that

$$p_i = \frac{1}{n} \cdot \frac{1}{1 + \nu_1(\omega_1(t_i) - 1) + \cdots + \nu_q(\omega_q(t_i) - 1)}, \tag{3.3}$$

which together with the constraints gives a set of equations

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{\omega_j(t_i) - 1}{1 + \nu_1(\omega_1(t_i) - 1) + \cdots + \nu_q(\omega_q(t_i) - 1)} = 0, \quad j = 1, \ldots, q. \tag{3.4}$$

Substitute p_i in $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G)$, the log-likelihood becomes up to a constant,

$$\begin{aligned} \ell &\equiv \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) \\ &= -\sum_{i=1}^{n} \log[1 + \nu_1(\omega_1(t_i) - 1) + \ldots + \nu_q(w_q(t_i) - 1)] \\ &\quad + \sum_{i=1}^{q} \sum_{j=1}^{n_i} (\alpha_i + \boldsymbol{\beta}_i' \boldsymbol{h}(x_{ij})). \end{aligned} \tag{3.5}$$

To get expressions for ν_j, we set $\partial l / \partial \alpha_j = 0$, $j = 1, \ldots, q$, and using equation (3.4), we obtain

$$\nu_j = \frac{n_j}{n}, \quad j = 1, \ldots, q.$$

Substituting these values of ν_j in equation (3.3), we have

$$p_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \omega_1(t_i) + \cdots + \rho_q \omega_q(t_i)}, \tag{3.6}$$

where $\rho_j = n_j/n_m$, $j = 1, \ldots, q$, and the value of the profile log-likelihood up to a constant as a function of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ only is

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{i=1}^{n} \log[1 + \rho_1 w_1(t_i) + \ldots + \rho_q w_q(t_i)] + \sum_{i=1}^{q} \sum_{j=1}^{n_i} (\alpha_i + \boldsymbol{\beta}_i' \boldsymbol{h}(x_{ij}))\,. \tag{3.7}$$

The term $\log[1 + \cdots]$ is due to the definition of the ρ_j and $\omega_j(t_i)$.

The score equation for $j = 1, \ldots, q$ are therefore,

$$\begin{cases} \frac{\partial l}{\partial \alpha_j} &= -\sum_{i=1}^{n} \frac{\rho_j \omega_j(t_i)}{1+\nu_1\omega_1(t_i)+\cdots+\nu_q\omega_q(t_i)} + n_j = 0 \\ \frac{\partial l}{\partial \boldsymbol{\beta}_j} &= -\sum_{i=1}^{n} \frac{\rho_j \boldsymbol{h}(t_i)\omega_j(t_i)}{1+\nu_1\omega_1(t_i)+\cdots+\nu_q\omega_q(t_i)} + \sum_{i=1}^{n_j} n_j = 0 \end{cases}.$$

Solving the above score equation, we obtain the maximum likelihood estimators of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and consequently by substitution also

$$\hat{p}_i = \frac{1}{n_m} \cdot \frac{1}{1 + \rho_1 \exp\{\hat{\alpha}_1 + \hat{\boldsymbol{\beta}}_1' \boldsymbol{h}(t_i)\} + \cdots + \rho_q \exp\{\hat{\alpha}_q + \hat{\boldsymbol{\beta}}_q' \boldsymbol{h}(t_i)\}} \tag{3.8}$$

therefore the maximum likelihood estimator of the reference distribution function $G(x)$ is obtained by summing over $\hat{p}$:

$$\hat{G}(x) = \frac{1}{n_m} \cdot \sum_{i=1}^{n} \frac{I(t_i \le x)}{1 + \rho_1 \exp\{\hat{\alpha}_1 + \hat{\boldsymbol{\beta}}_1' \boldsymbol{h}(t_i)\} + \cdots + \rho_q \exp\{\hat{\alpha}_q + \hat{\boldsymbol{\beta}}_q' \boldsymbol{h}(t_i)\}}\,. \tag{3.9}$$

It is argued in the Appendix that the estimators $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ are asymptotically normal as $n \to \infty$,

$$\sqrt{n} \begin{pmatrix} \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} \Rightarrow \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.10}$$

The vectors $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ denote the true parameters, and $\boldsymbol{\Sigma} = \boldsymbol{S^{-1}VS^{-1}}$, where the matrices $\boldsymbol{S, V}$ extend the results in Fokianos et al. (2001) [13] to a vector tilt and are also given in the Appendix. See Lu (2007) for a more detailed proof [40].

3.2.4 Hypothesis and Test Statistics

The null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ implies distributional homogeneity: $g_1(x) = g_2(x) = ... = g_m(x) \equiv g(x)$. We can use several test statistics to test this hypothesis. See Fokianos et al. (2001), Keziou and Leoni-Aubin (2005) and Fokianos (2006) for details [13, 27, 14].

χ_1 test statistic: Define a symmetric matrix $\boldsymbol{A}_{11}$ in terms of the relative sample sizes ρ_j,

$$\boldsymbol{A}_{11} = \begin{cases} \frac{\rho_j[1+\sum_{k\neq j}^{q}\rho_k]}{(1+\sum_{k=1}^{q}\rho_k)^2}, & \text{if } j = j' \\ \frac{-\rho_j\rho_{j'}}{(1+\sum_{k=1}^{q}\rho_k)^2}, & \text{if } j \neq j' \end{cases} \qquad j = 1, 2, \ldots, q \tag{3.11}$$

Then $\boldsymbol{A}_{11}$ is nonsingular. Under H_0, we deduce from the Appendix

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}'(t)] \\ \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}(t)] & \boldsymbol{A}_{11} \otimes E[\boldsymbol{h}(t)\boldsymbol{h}'(t)] \end{pmatrix}$$

and

$$\boldsymbol{V} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{11} \otimes Var[\boldsymbol{h}(t)] \end{pmatrix}$$

where $Var[\boldsymbol{h}(t)]$ is the covariance matrix of $\boldsymbol{h}(t)$ with respect to the reference distribution and all moments ($E[\boldsymbol{h}'(t)]$, $E[\boldsymbol{h}(t)\boldsymbol{h}'(t)]$) are evaluated with respect to the reference distribution. See Appendix for the details. The sub-matrices defining $\boldsymbol{S}$

and $\boldsymbol{V}$ have dimensions $q \times q, q \times qp, qp \times q, qp \times qp$, respectively. Consider the Wald-type statistic,

$$\chi_1 = n\hat{\boldsymbol{\beta}}'(\boldsymbol{A}_{11} \otimes Var[\boldsymbol{h}(t)])\hat{\boldsymbol{\beta}}. \tag{3.12}$$

This is an extension to a vector-valued $\boldsymbol{h}(t)$ of the χ_1 test statistic reported in Fokianos et al. (2001). It follows under H_0 that χ_1 is approximately distributed as χ^2 with qp degrees of freedom, and H_0 can be rejected for large values. Here $q = m - 1$ and p is the length of $\boldsymbol{\beta}_j$ which depends on the choice of the tilt function $\boldsymbol{h}$. For example, if $h(x) = x$, then $p = 1$, and if $\boldsymbol{h}(x) = (x, x^2)'$, then $p = 2$, and so on. The particular form of the χ_1 statistic (3.12) is due to the great simplification of $\boldsymbol{V}, \boldsymbol{S}$ under the hypothesis.

χ_2 test statistic: A general linear hypothesis $\boldsymbol{H\theta} = \mathbf{c}$ can be tested by means of

$$\chi_2 = n(\boldsymbol{H}\hat{\boldsymbol{\theta}} - \mathbf{c})'(\boldsymbol{H\Sigma H}')^{-1}(\boldsymbol{H}\hat{\boldsymbol{\theta}} - \mathbf{c}) \tag{3.13}$$

where $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_q, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_q')'$, $\boldsymbol{H}$ is $p' \times [(1+p)q)]$ predetermined matrix of rank p', $p' < (1+p)q$, $\mathbf{c}$ is a vector in $\Re^{p'}$, and the variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{S}^{-1}\boldsymbol{V}\boldsymbol{S}^{-1}$. It follows under H_0 that χ_2 is asymptotically distributed as χ^2 with (p') degrees of freedom provided the inverse exists (Sen and Singer, 1993, page 239) [56], and H_0 is rejected for large values.

Basically, the χ_1 test and the χ_2 test are both Wald type tests. The simulation results show that the χ_2 test is slightly more powerful than the χ_1 test. But the χ_1 test is easy to apply without inverting the $\boldsymbol{S}$ matrix, while the χ_2 test can be easily generalized to test any linear function of the parameter $\boldsymbol{\beta}$.

Likelihood ratio test statistic: A third possibility is to use the likelihood ratio test (LR-test),

$$\begin{aligned} LR &= -2[\ell(0,0) - \ell(\hat{\alpha}, \hat{\boldsymbol{\beta}})] \\ &= -2\sum_{i=1}^{n} \log[1 + \rho_1 \hat{w}_1(t_i) + \ldots + \rho_q \hat{w}_q(t_i)] \\ &\quad +2\sum_{i=1}^{q}\sum_{j=1}^{n_i}[\hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \boldsymbol{h}(x_{ij})] + 2n \log\left[1 + \sum_{i=1}^{q} \rho_i\right] \end{aligned} \tag{3.14}$$

Under H_0, LR is asymptotically approximately distributed as χ^2 with qp degrees of freedom, and H_0 is rejected for large values. In a few certain circumstances, this test is somewhat problematic since $(\boldsymbol{\alpha}, \boldsymbol{\beta}) = (\mathbf{0}, \mathbf{0})$ is a boundary point, an issue discussed rigorously in Keziou and Leoni-Aubin (2005) [27]. However, our experience indicates that in testing $\boldsymbol{\beta} = \mathbf{0}$, the LR-test works very well.

3.3 Semiparametric Cluster Detection

Since the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ means equal distributions, and homogenous distributions means no cluster in the cluster detection problem, the cluster detection problem becomes a special case for semiparametric density model with $m = 2$. Similar to Kulldorff's scanning procedure, the semiparametric cluster detection method applies the density ratio model to movable variable-size scanning window to scan the whole study region and performs for each window a two-sample test without assuming a specific probability model. The data can be either continuous or discrete. Since the significance comes from the χ^2-test, there is no need to do the time consuming Monte Carlo hypothesis testing procedure, hence it is not nec-

essary to know a priori the number of cases in the region. We use the same scanning procedure as Kulldorff's and select the primary cluster candidate corresponding to the largest test statistic or the smallest p-value (or q-value in the case of multiple testing) as the true cluster. The following is an illustration of semiparametric cluster detection.

Consider the 5×5 region consisting of 25 cells shown in Figure 3.1(a). Within each cell, numbered from 1 to 25, there are hundreds of binary observations generated randomly. The rate in one of the cells is higher than the rate in the rest of the region. This is the true cluster to be detected. We applied to these simulated data both Kulldorff's scan statistic method with the Bernoulli model and the semiparametric density ratio method with scalar $h(x) = x$. Starting with the first cell, the window size varied from a size roughly as large as a cell size to no more than 50% of the whole study region. This was repeated for each cell. Figure 3.1 (b)-(e) shows some snapshots of the intermediate stages during scanning of the whole study region. Both methods detected correctly the true cluster shown in Figure 3.1(f).

In addition, for the case of $m = 2$, the semiparametric χ_1 test and likelihood ratio test can be simplified to a simpler form as follows. A one-sided test can also be obtained from χ_1 test when $h(x)$ is scalar function.

χ_1 test statistic:

$$\chi_1 \equiv n\,\hat{\boldsymbol{\beta}}_1' \left(\frac{\rho_1}{(1+\rho_1)^2}\, Var[\boldsymbol{h}(t)] \right) \hat{\boldsymbol{\beta}}_1 \tag{3.15}$$

where $\rho_1 = n_1/n_2$ and $Var[\boldsymbol{h}(t)]$ is the covariance matrix of $\boldsymbol{h}(t)$ with respect to the reference distribution. It follows under H_0 that χ_1 is approximately distributed

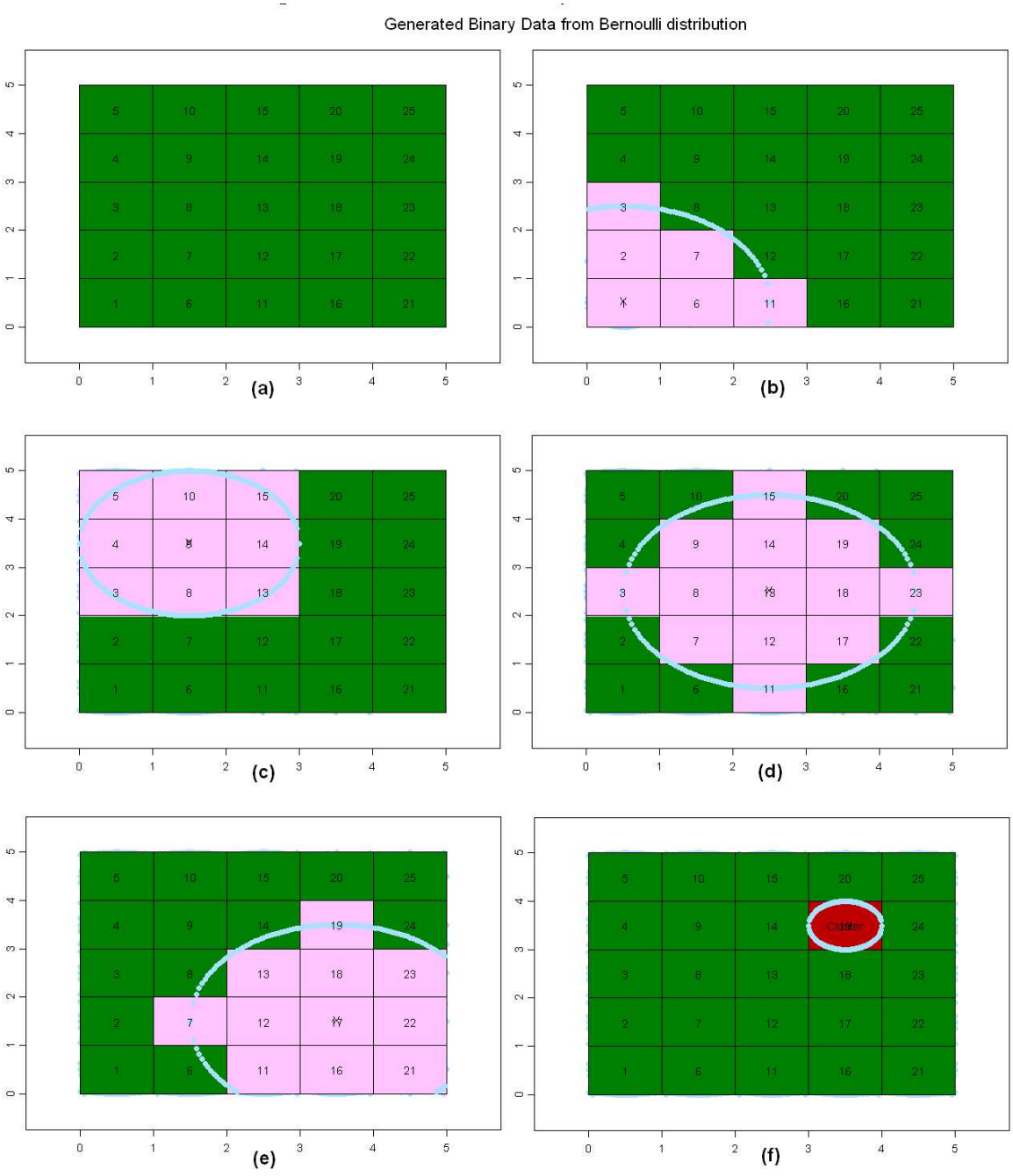


Figure 3.1: (a) The whole study region. (b,c,d,e) Intermediate stages during the scan. (f) The red region is the true cluster. The true cluster was detected by both methods.

as χ^2 with p degrees of freedom, where p is the dimension of $\boldsymbol{\beta}_1$, which is the same as that of the function $\boldsymbol{h}(x)$. The null hypothesis $H_0 : \boldsymbol{\beta}_1 = 0$ is rejected for large values of χ_1. In practice, $Var[\boldsymbol{h}(t)]$ is replaced by its estimator [13, 26]. Moreover, if the data are 0-1 binary data, the χ_1 statistic takes the direct form of equation (A.21) in the Appendix.

Likelihood ratio test statistic:

$$\begin{aligned} LR \;\equiv\; & -2\sum_{i=1}^{n}\log[1+\rho_1\cdot\exp\{\alpha_1+\boldsymbol{\beta}_1'\boldsymbol{h}(t_i)\}]+2\sum_{j=1}^{n_1}[\hat{\alpha}_1+\hat{\boldsymbol{\beta}}_1'\boldsymbol{h}(x_{1j})] \\ & +2n\log(1+\rho_1) \end{aligned} \tag{3.16}$$

Under $H_0 : \boldsymbol{\beta}_1 = 0$, LR is asymptotically approximately distributed as χ^2 with p degrees of freedom, and H_0 is rejected for large values. Recall that p is the dimension of $\boldsymbol{\beta}_1$ and it depends on the choice of the function $\boldsymbol{h}(x)$. Similarly, when the data are 0-1 binary, the likelihood ratio statistic reduces to equation (A.22) in the Appendix. Interestingly, this semiparametric likelihood ratio statistic is equivalent to Kulldorff's scan statistic under the Bernoulli model [70].

One-sided test statistic: The χ_1, χ_2 and LR test statistics are all two-sided tests of $H_0 : \boldsymbol{\beta} = 0$. When $m = 2$ and $h(x)$ takes a scalar form, such as $h(x) = x$ or $h(x) = \log x$, the parameter β becomes a scalar, so it is possible to derive a one-sided test as in equation (3.17):

$$Z_{1-sided} = \frac{\hat{\beta}-\beta_0}{\sigma_\beta} \rightarrow N(0,1) \tag{3.17}$$

where $\beta_0 = 0$ and σ_β^2 is the variance of $\hat{\beta}$ obtained from the covariance matrix $\boldsymbol{\Sigma}$. In practice, under H_0, this one-sided test statistic can be obtained by the square root

of the χ_1 test statistic, keeping the same sign as $\hat{\beta} - \beta_0$.

The semiparametric approach has several advantages as follows:

① The reference (or background) distribution, $G(x)$, and all the parameters such as $\boldsymbol{\beta}_1$ are estimated from the *combined data* $\boldsymbol{t}$, not just from a single sample either inside the window or outside the window.

② For a properly chosen $\boldsymbol{h}(x)$, the above tests are quite powerful. Gagnon (2005) shows that for $m = 2$ the χ_1-test can be more powerful than the common t-test for a known $\boldsymbol{h}(x)$ but unspecified distributions [16]. Moreover, simulation results indicate that the χ_1-test competes well with the corresponding F-test (Fokianos et al., 2001) [13].

③ In testing equidistribution within exponential families, other than an assumption regarding the tilt function $\boldsymbol{h}(t)$, the semiparametric density ratio method does not require specific distributional assumptions.

④ The semiparametric method can be applied to either continuous or discrete distributions.

⑤ Assuming sufficient large samples, since the asymptotic distributions of the above mentioned test statistics are known, in principle there is no need for the time consuming Monte-Carlo methods to compute the p-values. In case of small sample size, then we can still bear with Monte-Carlo methods to get p-values.

Since each scan window is associated with a semiparametric statistic during scanning, the method results in a large number of tests and test statistics. To alleviate this multiple testing problem and reduce the possibility of false significance, a control of false discovery rate (FDR) procedure is employed. The following section gives a brief description of the FDR methodology and Storey's q-value method we used in this paper.

3.4 FDR Method and q-value

To account for the multiple-testing problem induced by the large set of overlapped scanning windows, we use Storey's false discovery rate (FDR) method to derive the significance of the detected cluster candidate. This FDR method replaces the original p-value of each scan window by a q-value. We briefly describe in this section the FDR methodology as well as the q-value method used in this research.

Controlling the false discovery rate (FDR) is a less conservative way to handle multiple testing problems. It was first proposed by Benjamini and Hochberg (1995) [3]. Since then, the FDR methodology has been further developed and applied in many fields, especially in genomic research [66, 54]. FDR is defined to be the expected proportion of falsely rejected hypotheses (false positives) as in equation (3.18):

$$FDR = E\left[\frac{V}{max(1,R)}\right] = E\left[\frac{V}{R}|R>0\right]\cdot Pr(R>0). \tag{3.18}$$

where V and R are defined in Table 3.1. From the table it is clear that if $m = m_0$, then all the null hypotheses are true, and FDR is equivalent to the family-wise

error rate (FWER). To see that, recalling that FWER is defined as $P(V \geq 1)$, $m = m_0$ makes $S = 0$, and hence $E[V/R|R > 0] = 1$ for all $R > 0$, and therefore $FDR = 1 * P(R > 0) = P(V \geq 1) = FWER$. If $m_0 < m$, then $FDR \leq FWER$, which means a potential gain in power at the cost of increasing the likelihood of making type I errors [3].

Table 3.1: Classification of m hypothesis tests

Hypothesis	***# Accepted***	***# Rejected***	***Total***
# of true null Hypotheses	U	V	m_0
# of true alternatives	T	S	m_1
Total	W	R	m

Storey (2002) and Storey et al. (2004) improved the original Benjamini and Hochberg FDR methodology by estimating $\pi_0 = m_0/m$, the proportion of true null hypotheses [60, 63]. In addition, Storey used *pFDR* as in equation (3.19) instead of FDR (3.18):

$$pFDR = E\left[\frac{V}{R}|R > 0\right] \tag{3.19}$$

In many cases, when m, the total number of hypotheses, is large, there always are significant ones, which makes $Pr(R > 0) \approx 1$. Thus, pFDR (eq. 3.19) is close to FDR (eq. 3.18) in numerical value, but pFDR has some conceptual advantages. For instance, when the rejection region is smaller, namely α-level goes to 0, the quantity of FDR goes to 0 as $Pr(R > 0)$ goes to 0. It doesn't mean the actual chance of false positive decreases to 0. However the quantity of pFDR goes to the

π_0, which is what we would expected. Since when the rejection region is smaller until only one p-value falls into the region, without any information about the null of alternative hypotheses, it makes sense to use π_0, the proportion of the nulls among all hypotheses, as the estimation of the chance of the false positive. A thorough motivation of using pFDR rather than FDR can be found in Storey (2003) [62].

Recall that the p-value gives an error measurement of an observed statistic with respect to type-I error. In a general setting, the p-value of an observed statistic $T = t$ is defined to be

$$p\text{-value}(t) = \min_{\Gamma_\alpha:\, t \in \Gamma_\alpha} \{Pr(T \in \Gamma_\alpha | H = 0)\}$$

where $H = 0$ means under the null hypothesis and Γ_α is a nested rejection region parameterized with α and, for $\alpha \leq \alpha'$, $\Gamma_\alpha \subseteq \Gamma_{\alpha'}$ holds.

The q-value is defined to be the pFDR analogue of the p-value. It gives the error measurement with respect to pFDR for each observed test statistic of each particular hypothesis. More precisely, the q-value of one particular observed test statistic $T = t$ from a set of tests can be defined to be

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha:\, t \in \Gamma_\alpha\}} \{pFDR(\Gamma_\alpha)\}. \tag{3.20}$$

In this way, the q-value is the minimum pFDR that can occur when rejecting a statistic with value t for the set of nested rejection regions. In addition, for a set of hypothesis tests conducted with independent p-values, the q-value of the corresponding observed p-value can be simplified to

$$q\text{-value}(t) = \inf_{\gamma \geq p} \{pFDR(\gamma)\} = \inf_{\gamma \geq p} \{\frac{\pi_0 \gamma}{Pr(P \leq \gamma)}\} \tag{3.21}$$

where the rejection region is denoted by the p-value interval $[0, \gamma]$ for some $\gamma \geq 0$ instead of the more abstract rejection regions Γ. Storey J.D. and Tibshirani R. (2001, 2003) shows that the q-value method holds similar properties under either independence or dependence cases [59, 61].

In our semiparametric scan situation, we generate a lot of overlapping scan windows, and each window associates to a hypothesis test and a test statistic, so m here is usually large. Because of the large m, we adopted the algorithm in Storey and Tibshirani (2003) [61] to estimated the q-value for each scan window as follows:

1. Obtain the p-value for each scan window, and sort:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}.$$

2. Estimate π_0 using a cubic spline function. First, for a range of λ, say $\lambda = 0, 0.01, \ldots, 0.95$, calculate

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_{(j)} \geq \lambda\}}{m(1-\lambda)} \text{ for each } \lambda.$$

   Then let $\hat{f}(\lambda)$ be the natural cubic spline fit to $\hat{\pi}_0(\lambda)$. Finally, set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$.

3. Calculate

$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \left(\frac{\hat{\pi}_0 \, m \, t}{\#\{p_{(j)} \leq t)\}} \right) = \hat{\pi}_0 \, p_{(m)}, \text{ where } 0 < t < 1.$$

4. For $i = m-1, m-2, \ldots, 1$, calculate the estimated q-value for the ith most significant one as

$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \left(\frac{\hat{\pi}_0 \, m \, t}{\#\{p_j \leq t)\}} \right) = \min \left(\frac{\hat{\pi}_0 \, m \, p_{(i)}}{i}, \; \hat{q}(p_{(i+1)}) \right).$$

5. Choose the region with the largest test statistic, for example, the largest likelihood ratio test statistic, as the primary cluster candidate, and its q-value is $q(p_{(1)})$, the smallest q-value among all the tests. If $q(p_{(1)})$ is less than a predecided false discovery rate, say $q(p_{(1)}) < 0.05$, we claim there is clustering and the located primary candidate is a true cluster region.

Chapter 4

Power Study

In this chapter, we study the power of the semiparametric cluster detection method compared with Kulldorff's method through Monte Carlo simulations. One type of study is a limited power study where the location and the size of the clusters are already known without scanning. This limited power study generated power curves for semiparametric method, Kulldorff's method, and the focused test for data under different probability distributions. Another type of study is a comprehensive power study comparing the semiparametric method and Kulldorff's method without knowing any information about the clusters. In this study one must scan the whole study region to locate the cluster candidate. Since both Kulldorff's and the semiparametric scan statistics methods are suitable for both binary and non-binary data, we use both types of data to perform the comprehensive power study.

Both the limited and comprehensive power studies show that for binary data, the semiparametric cluster detection method and its competitor, Kulldorff's celebrated scan statistics method, both achieve similar high power in detecting unknown hot-spot clusters. When the data are not binary, the semiparametric methodology is still applicable, but Kulldorff's method may not be as it requires the choice of a correct probability model, namely the correct scan statistic, in order to achieve power comparable to that achieved by the semiparametric method. Kulldorff's method

with an inappropriate probability model may lose power.

4.1 Limited Power Study

The main purpose of the limited power study is to compare the performance of the semiparametric and Kulldorff's method in determining the significance of a known cluster candidate. This can be criticized on the grounds that we do not scan the area for clusters without the prior knowledge of where those clusters are located. When the cluster location is known, it is more appropriate to compare the semiparametric method with what is known as *focused tests* described in Waller and Lawson (1995) and Lawson et al. (1999) [68, 38]. So we first briefly introduce the Lawson-Waller focused test in the following subsection.

4.1.1 Focused Tests

Focused tests detect clusters with increased risk of disease relative to a source of exposure or *focus*. Since the problem of multiple testing in cluster detection is avoided, focused tests tend to have higher power than cluster detection tests.

A well known focused test is the Lawson-Waller score test applied in disease surveillance [68]. Accordingly, the study area is divided into I subregions where the population size in subregion i is n_i, $i = 1, \ldots, I$. Denote the number of cases in region i by C_i. The null hypothesis is that the C_i are independent Poisson with mean $E(C_i) = \lambda n_i$, $i = 1, \ldots, I$, against the (hot spot) alternative

$$H_1 : E(C_i) = \lambda n_i(1 + g_i\epsilon), \quad i = 1, \ldots, I \tag{4.1}$$

where g_i is a measure of exposure to a focus, and $\epsilon > 0$ controls the increase in risk. A reasonable test statistic is the Lawson-Waller score statistic

$$U = \sum_{i=1}^{I} g_i \cdot (C_i - E(C_i)) \tag{4.2}$$

which under the null hypothesis has mean 0 and variance $\sum_{i=1}^{I} g_i^2 \lambda n_i$, and $U/\sqrt{Var(U)}$ is asymptotically standard normal. This statistic can be used in testing for trend in Poisson random variables.

4.1.2 Data and Simulation plan

We consider two regions, A and B, where A consists of 100 subregions or cells, and B of 1000 cells (except that in the Bernoulli case below the number of cells were 200 and 5000, respectively). The population size in every cell is identical. The smaller region can be thought of as a cluster candidate, whereas the larger region could represent the rest of the area, or some reference or baseline region. The incidence rates in every cell in A, represented by either a case probability, mean, or occurrence rate, are identical. The same holds for B. Thus, in terms of occurrence rate, it is λ_A in every cell in A, and it is λ_B in every cell in B. Independent count data were generated in an identical manner in every cell in A, one count observation per cell, and likewise, independent count data were generated in the same way in every cell in B, a single count observation per cell. In the Bernoulli case every cell contains either 0 or 1. The parameters for B never change, but those for A change relative to B.

In this way two samples were generated repeatedly from "within the window" (from A) and "outside the window" (from B), respectively, with the same or different parameters, as needed. In our study, the sample size "within the window" is smaller than the sample size "outside the window" since in practice the size of a true cluster tends to be small relative to the whole study region.

In this setup, the Lawson-Waller focused test assumes independent Poisson cell counts with $\lambda_A = \lambda_B$ under H_0, versus $\lambda_A = \lambda_B(1 + g_i\varepsilon)$ under H_1, and we let

$$g_i = \begin{cases} 1 & \text{if cell } i \text{ is in } A \\ 0 & \text{if cell } i \text{ is in } B \end{cases}$$

When β is a scalar, we use the Z test statistic of equation (3.17) to compare the detecting power with Kulldorff's scan statistic. When $\boldsymbol{\beta}$ is not a scalar, we use the χ_1, χ_2 and LR test statistics in two-sided tests, and adjust the original Kulldorff test into a two-sided test. Similar remarks hold for focused tests. In this way we compare one-sided with one-sided and two-sided with two-sided tests.

The following series of figures shows the results of the power simulation. Each power curve was obtained from 300 runs, and the size of all the tests was controlled at the same level of 5%. For Kulldorff's Monte-Carlo hypothesis test we used 10000 replications.

4.1.3 Results for Various Probability Distributions

Bernoulli: Figure 4.1 shows the estimated power curves in the Bernoulli case for one-sided tests. The null probability is $p_0 = 0.03$ and p ranges in the interval $[0.03, 0.13]$. Kulldorff's method is applied under the assumption the data are Bernoulli, whereas the semiparametric method is applied with $h(x) = x$. Evidently, the Lawson-Waller score test, designed for count data, dominates both Kulldorff's and the semiparametric tests, and the latter two exhibit very close power curves.

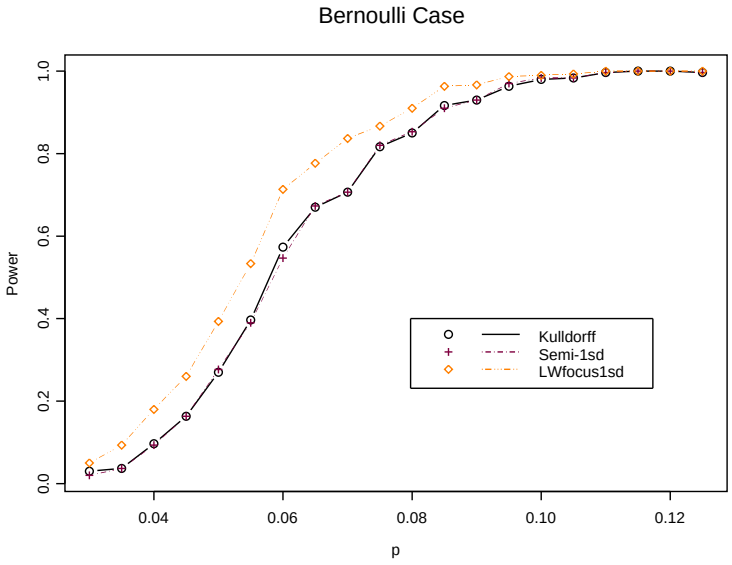


Figure 4.1: Power curves for one-sided tests in the Bernoulli case. Scalar β, $h(x) = x$. The focused test dominates the two other tests.

Poisson: Figure 4.2 shows the power curves for one-sided tests and Poisson data. The Poisson parameter ranges from the null intensity of 5.0 to 6.5. Kulldorff's method is applied under a Poisson model, and the semiparametric method is applied with $h(x) = x$. The power curves of the semiparametric and focused tests are fairly close, and both dominate Kulldorff's.

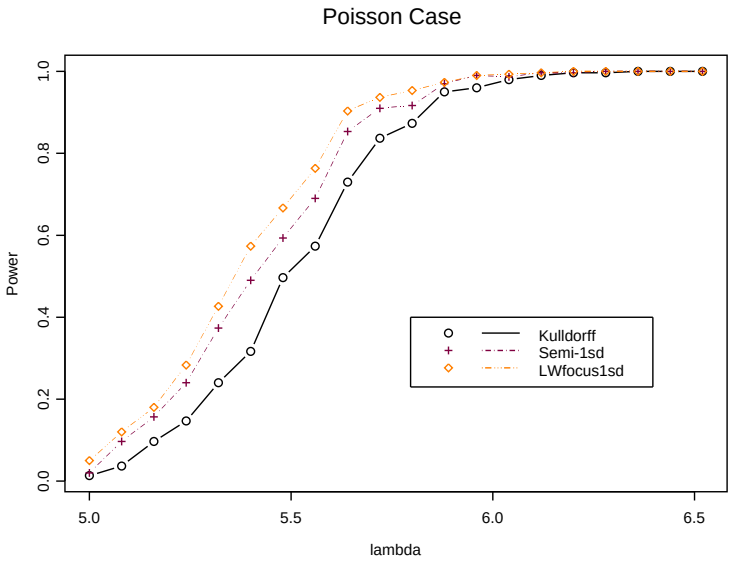


Figure 4.2: Power curves for one-sided tests in the Poisson case. Scalar β, $h(x) = x$. The power curves from the semiparametric and focused tests are fairly close.

Clipped Poisson: Figure 4.3 shows the power curves for one-sided tests and data generated from clipped Poisson observations. The parameters are the same as in the previous Poisson case. Equation 4.3 describes the clipping operation,

$$z = \begin{cases} 2 & (x <= 2) \\ x & (2 < x <= 10) \\ 10 & (x > 10) \end{cases} \tag{4.3}$$

Kulldorff's method is applied under the Poisson model, and the semiparametric method still uses $h(x) = x$. The semiparametric method gives relatively higher power, apparently due to hard limiting.

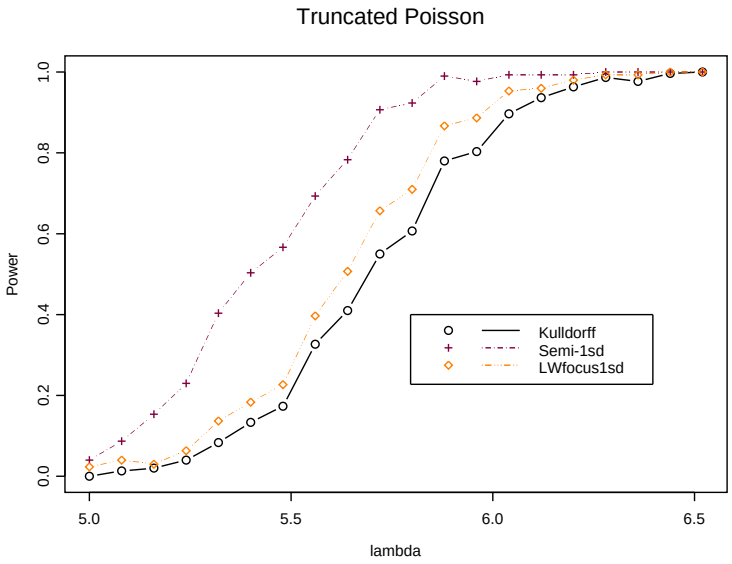


Figure 4.3: Power curves for one-sided tests in a clipped Poisson case. Scalar β, $h(x) = x$. The semiparametric method gives relatively higher power.

Quantized Normal I : In this and the next two examples we turn to count data generated by quantizing normal observations. This is motivated by real situations when the data are non-Poisson count data, but not knowing the true distribution the Poisson assumption is made nonetheless. In the present case the Poisson assumption is sensible up to a point as our simulation shows. The semiparametric method obviates this assumption.

The quantized data were obtained from the integer part of the original normal data. The original normal samples share the same variance ($\sigma^2 = 16$), but the mean μ of the A samples ranges from the null $\mu_0 = 9$ to $\mu = 12$. Kulldorff's method is applied under the Poisson model, the semiparametric method uses $h(x) = x$, and the tests are one-sided. Figure 4.4 shows the resulting power curves. The focused test dominates both Kulldorff's and the semiparametric tests, and the last two perform very similarly. However, from Figure 4.5, the situation changes dramatically for the same variance 16 but much higher means ranging from 50 to 53. This time the semiparametric test clearly dominates the two other tests. The situation here resembles that of the doubly truncated Poisson case depicted in Figure 4.3 since the quantized data stay away from very small and very large values with a high probability.

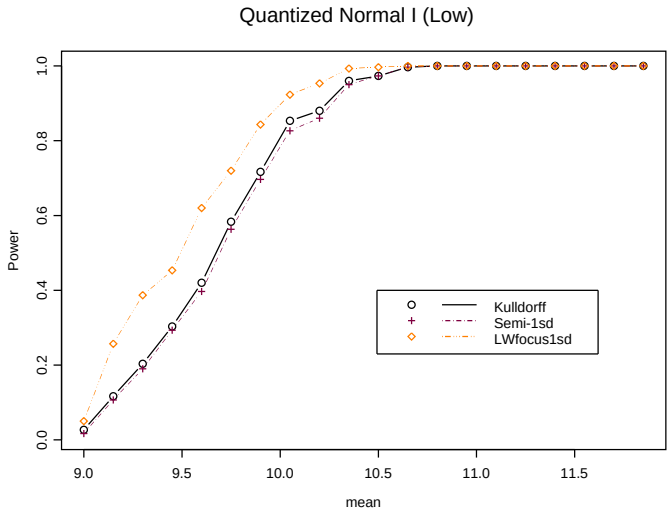


Figure 4.4: Power curves for one-sided tests applied to Quantized normal samples with the same variance 16 but different means. Scalar β, $h(x) = x$. The focused test gives higher power.

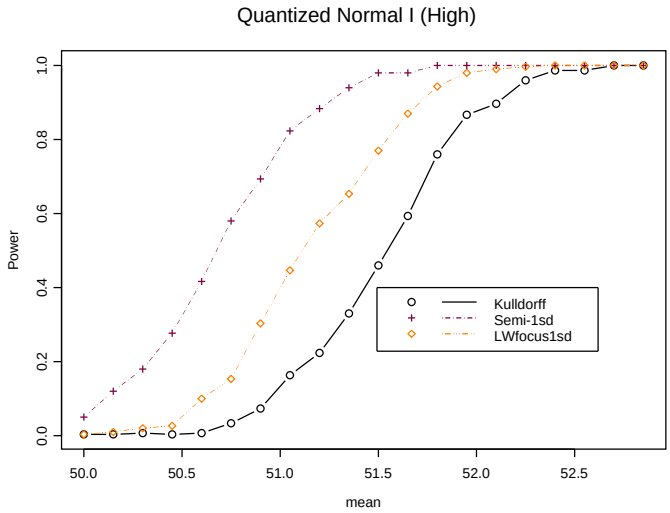


Figure 4.5: Power curves for one-sided tests applied to Quantized normal samples with the same variance 16 but different relatively high means. Scalar β, $h(x) = x$. The semiparametric test clearly dominates the other two tests.

Quantized Normal II : Figure 4.6 shows the power resulting from two-sided tests applied to integer quantized normal data as in the previous example. The quantized samples are derived from normal data with the same mean $\mu = 13$ but different variances, respectively, where the variance ranges from 4 (null) to 10. Kulldorff's method is applied under the Poisson model, and the semiparametric method uses $h(x) = (x, x^2)'$, a model suggested by the normal distribution. In this case the three semiparametric tests are much more powerful than the other two tests whose power is almost identical. Kulldorff's method with the Poisson model and the focused count model seem not appropriate for the present case.

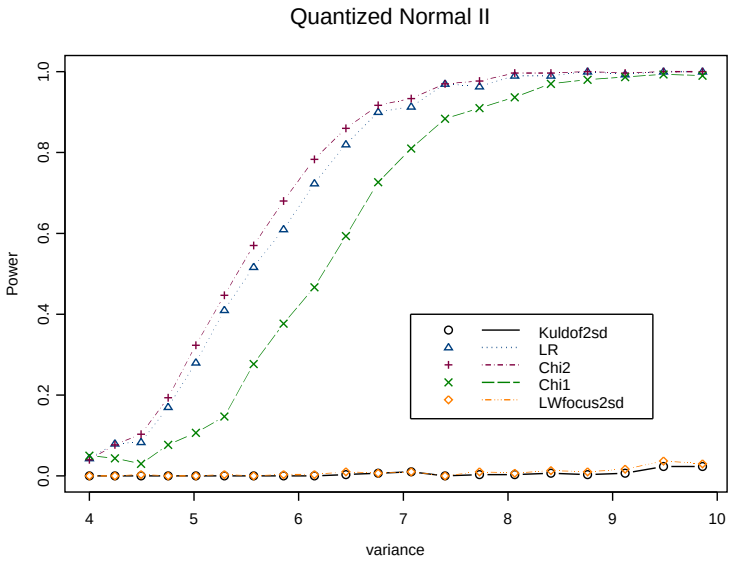


Figure 4.6: Power curves for two-sided tests applied to quantized normal samples with the same mean but different variances. The semiparametric method uses $h(x) = (x, x^2)'$, χ_1, χ_2, LR. The semiparametric tests markedly dominate the two other tests.

Quantized Normal III : This case is the same as the previous one except that both the means and the variances are different. Recall that here and elsewhere, the null hypothesis is equidistribution. The mean ranges from the null of $\mu = 20$ to $\mu = 21$, and the corresponding variance is $\sigma^2 = 4$ when $\mu = 20$, and is $\sigma^2 = 7$ otherwise. From Figure 4.7 we see again that the three semiparametric tests are much more powerful than Kulldorff's and the focused test. This and the previous examples give an indication that Kulldorff's test and the Lawson-Waller focused test may not be suitable for integer samples with substantially different variances.

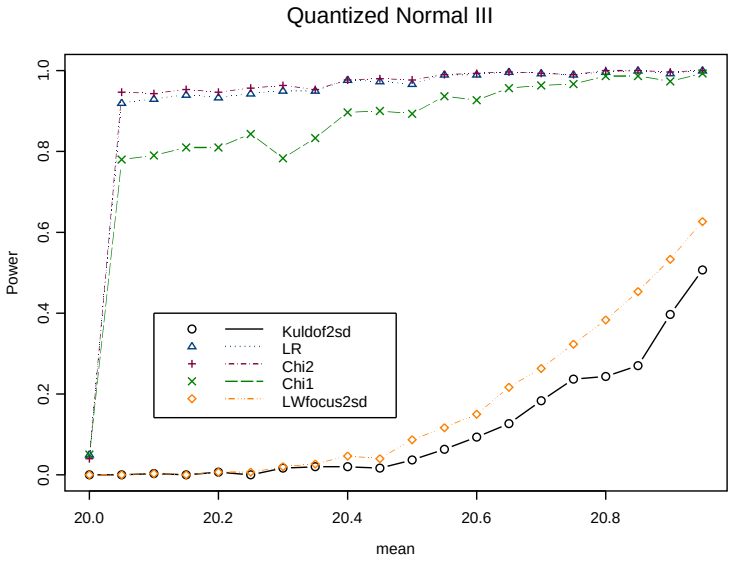


Figure 4.7: Quantized normal case as in Figure 4.6 but with different means and different variances. The semiparametric tests clearly dominate the two other tests.

From the above power results we find that Kulldorff's method and the Lawson-Waller focused test perform well for some types of count data but may lose power for count data with non-homogeneous variance, as well as count data which are far from

being Poisson. In contrast, the semiparametric method seems to perform relatively well under very different settings, without specifying any distribution except for the choice of the tilt function. In particular, our simulation results indicate that this semiparametric method is potentially useful across different types of data with changing regional means and/or variances. With a properly chosen tilt function $\boldsymbol{h}(x)$, the method can detect changes in both the mean and the variance.

4.2 Comprehensive Power Study: Overview

We compared the power of Kulldorff's and the semiparametric scan statistics methods in detecting potential clusters. Because there are various cluster patterns, with a single or multiple clusters, each cluster region may contain one or more counties or states. For simplicity, in our power comparison, exact accuracy is not required, we focus more on the existence instead of precise delineation of the cluster region . For instance, for a data set with a pattern of multiple cluster regions and multiple counties in each cluster region, we deem the detection successful whenever a significant q-value is obtained. We do not strictly require the detected cluster region to be exactly the same as originally simulated. The detected cluster region could fully or only partially cover the desired area.

4.3 Comprehensive Power Study: Binary Data

4.3.1 Binary Data Set and Simulation Plan

For binary data scans, we use the Northeastern U.S.A. purely spatial benchmark data consisting of 245 counties in northeastern United States, from Maine, New York, Rhode Island, Pennsylvania, Maryland, Washington DC, among others (Kulldorff et al., 2003) [32]. Each county is graphically represented by its centroid coordinates. The case data, the numbers of people who have breast cancer, are aggregated to county level with the total number of cases in northeastern states being fixed. The population data of each county are based on the female population of the 1990 census. The benchmark data set contains two types of data, hot-spot clusters and global clustering data. Figure 4.8 is a map of the northeastern U.S. states. The following briefly describes the simulated data and data sets. See Kulldorff et al. (2003) for details [32].

Hot-spot clusters: Data are generated by a first-order clustering model, where cases are located independently of each other and the relative risk is different in different geographical areas. In this US northeastern states benchmark data set, the cluster region can be either a single region containing one or more counties, or a collection of multiple regions, where the risk of breast cancer is much higher than in the rest of the area. Three types of cluster, rural, urban, and mixed, are generated depending on the location of the cluster. A rural cluster is a region which has a small population relative to a large graphical area, such as Grand Isle County

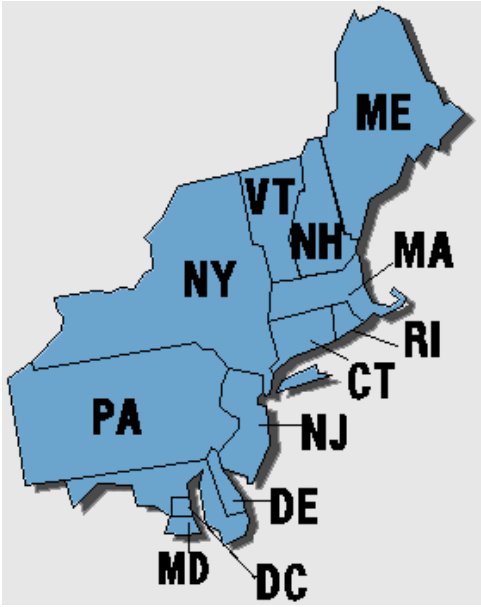


Figure 4.8: sketch map of the US northeastern states.

in northern Vermont close to the Canadian border. An urban cluster is a region which has a large population relative to a small graphical area, such as New York County which includes Manhattan. A mixed cluster is a region where a big city is surrounded by rural areas, such as Allegheny county in western Pennsylvania where Pittsburgh is located.

Global clustering: Data are generated by purely second-order clustering model, where any one particular case is randomly located, so that the relative risk is constant throughout the whole study region, but the location of cases are dependent on each other. Thus, under the alternative hypothesis of global clustering, cases are clustered wherever they occur in the region. In this benchmark data set, a certain number of cases are first generated to be randomly located throughout the whole northeastern states. These original cases then generate other new cases close

by. If each original case generates one additional case, we call them twins; if two additional cases are generated, we call them triplets. The case generation is based on a global chain rN-nearest neighbor rule. The global chain is constructed by a Hamiltonian cycle chain which passes through as many counties as possible exactly once, and any two counties next to each other on the chain always border each other graphically. For twins, the additional case is assigned to county j if $\sum_k I(d_{ik} < d_{ij})n_k < rN \leq \sum_k I(d_{ik} \leq d_{ij})n_k$, where n_k is the population size of county k, $N = \sum_k n_k$ is the total population size, r is some constant in the interval $(0, 0.5)$, and d_{ij} is the distance in one particular direction along the chain connecting county i and county j. For triplets, the two new cases are assigned in opposite directions along the chain. Data sets corresponding to different r were generated, where r is either deterministic or randomly selected from a probability distribution.

Notice that although the first- and the second-order clustering models are very different in generating the cases, the resulting point patterns may look quite similar, and hence indistinguishable.

This benchmark data set includes two groups of data with a total of 600 and 6000 simulated cases, respectively, for both hot-spot and global clustering data sets. The same null hypothesis of no cluster is used throughout where the relative risk for each county is equal, and the cases as well as their locations are independent of each other. In order to perform power comparison, 100000 random data sets with a total of 600 and 6000 cases were generated under the null hypothesis, respectively. These are used to estimate the critical cut-off point of significance. For each alternative

hypothesis of clusters which are called *scenarios* in this paper, 10000 random data sets were generated to estimate the power using the previous determined cut-off points.

For each group of fixed total cases, Kulldorff generated 35 hot-spot clustering scenarios and 26 global clustering scenarios for his power comparison. For instance, a scenario of "rural and urban 600, size 4" means a total of 600 cases were generated under the alternative hypothesis that the study region has two hot-spot clusters. One cluster is in a rural region including four counties, and the other one is in an urban region including four counties as well. A scenario of "global clustering twin 6000, exponential 0.02" means a total of 6000 cases were generated under the alternative hypothesis of global clustering. The value r is randomly generated from an exponential distribution with parameter 0.02.

In this paper, we did not use all the scenarios in the Northeastern US benchmark data. Instead, we randomly chose one or two scenarios from each clustering pattern. Finally, 9 hot-spot clustering scenarios and 6 global clustering scenarios are used in our power study for the binary data. After selecting these scenarios, the same data sets in each scenario were used for Kulldorff's method with the Poisson model (the Bernoulli model is also appropriate) and the semiparametric method was applied with the tilt function $h(x) = x$. In addition, the simplified semiparametric likelihood ratio test statistic (eq. 3.16) is used to accelerate the computation. All the tests are two-sided to detect for either high or low valued clusters.

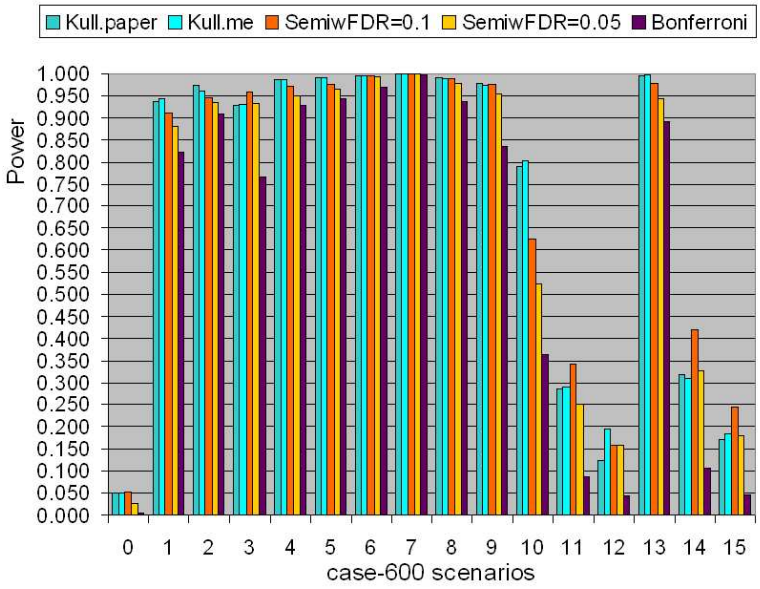
4.3.2 Results for Binary Data

The results of power comparison for the binary case-population data using the northeastern US benchmark data set are shown in Figures 4.9 and 4.10 for a total number of 600 and 6000 cases, respectively. All tests are two-sided tests. In each figure, scenario 0 is under the null hypothesis of no cluster, scenarios 1 to 9 are hot-spot clustering scenarios, and scenarios 10 to 15 are global clustering scenarios. Each scenario contains five quantities. They are "Kull.paper", "Kull.me", "SemiwFDR=0.1", "SemiwFDR=0.05", and "Bonferroni". "Kull.paper" is the corresponding power copied from Kulldorff et al. (2003) paper. "Kull.me" is the corresponding power computed by us based on Kulldorff Poisson model. The purpose of including "Kull.paper" here is to make sure our programming and computation are correct. If we are correct, the results of "Kull.me" should be similar to those in "Kull.paper". From Figures 4.9 and 4.10, we can see that they are almost equal, which confirm the validity of our computation. So in the latter of this section, we will use "Kulldorff's method" without distinguishing these two power results. "SemiwFDR=0.1" is the corresponding power computed based on a q-value significance level 0.1. "SemiwFDR=0.0.5" is the corresponding power computed based on a q-value significance level 0.05. The smaller the q-value significance level is, the harder it is to reject the null hypothesis, hence the lower the power to detect the cluster. "Bonferroni" is the corresponding power computed based on Bonferroni correction with the family-wise error rate 5%. Because Bonferroni correction is a popular but a conservative approach to handle multiple testing problems, we also

included it here to compare with the FDR methodology. We expect "Bonferroni" to have the lowest power among the five.

For scenario 0, both Figures 4.9 and 4.10 show that the type I errors of Kulldorff's method are all exactly 0.05 for both 600 cases and 6000 cases. This is expected since Kulldorff's method uses a Monte Carlo procedure to derive the cut-off point for the corresponding significance level. Since the significance level for Kulldorff's method in our power study is 0.05, the power computed from Kulldorff's method under the null hypothesis, which is the type I error, must be 0.05. The power from semiparametric method under the q-value significance level of 0.1 is 0.052 for the 600 cases, and 0.054 for the 6000 cases. It means that, if we allow a higher false discovery rate such as 10%, which is in favor of the alternative, the type I error of semiparametric method is slightly higher than Kulldorff's. If the q-value significance level is chosen as 0.05, the type I error of the semiparametric method reduced to 0.027 for both cases. This is also expected since lower q-value significance level works in favor of the null hypothesis. It shows that if one wants to make the power comparison under exactly the same type I error level, say 0.05, the q-value significance level must be set in the interval between 0.05 to 0.1. In addition, the type I error of the semiparametric method with Bonferroni correction is the lowest, which is not a surprise since Bonferroni correction is the most conservative.

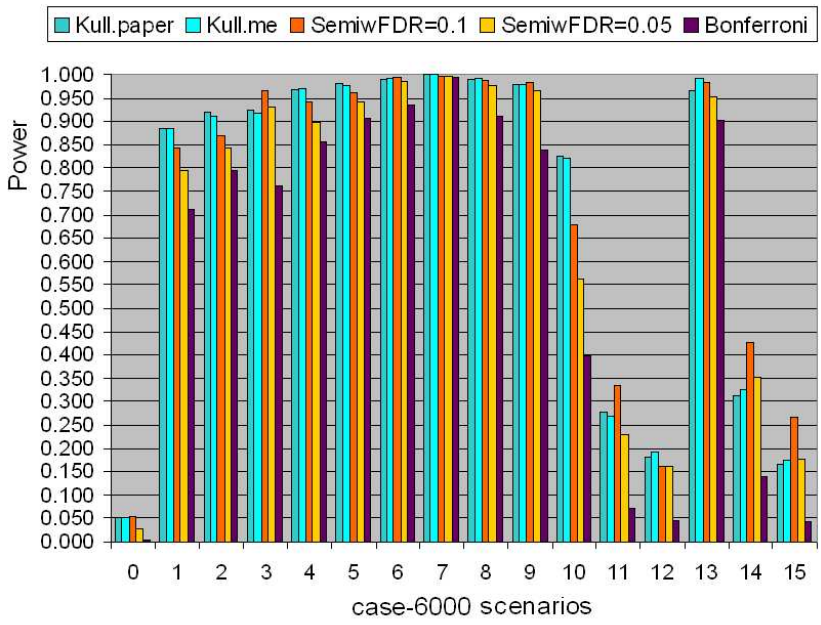
For scenarios No. 1 to 9, both figures show that the two methods work very well in detecting hot-spot clusters, but Kulldorff's method seems to be slightly more powerful. When the study area has a stronger pattern of clustering, such



No.	Kull.paper	Kull.me	SemiwFDR=0.1	SemiwFDR=0.05	Bonferroni
0	0.050	0.050	0.052	0.027	0.004
1	0.936	0.942	0.911	0.881	0.822
2	0.973	0.962	0.945	0.933	0.908
3	0.926	0.929	0.959	0.931	0.765
4	0.987	0.987	0.970	0.949	0.926
5	0.992	0.990	0.975	0.965	0.944
6	0.996	0.996	0.996	0.993	0.968
7	1.000	1.000	1.000	1.000	0.998
8	0.992	0.989	0.989	0.977	0.935
9	0.977	0.973	0.976	0.955	0.836
10	0.791	0.803	0.626	0.523	0.362
11	0.285	0.289	0.343	0.252	0.086
12	0.124	0.195	0.158	0.158	0.043
13	0.995	0.998	0.978	0.942	0.890
14	0.318	0.310	0.420	0.326	0.107
15	0.171	0.186	0.244	0.181	0.045

No.	Scenario
0	No cluster 600, under null hypothesis (type I error)
1	single cluster, mixed 600, size 1
2	single cluster, rural 600, size 4
3	single cluster, urban 600, size 16
4	two clusters, mixed and urban 600, size 1
5	two clusters, rural and urban 600, size 4
6	two clusters, rural and mixed 600, size 16
7	three clusters, rural, mixed and urban 600, size 1
8	three clusters, rural, mixed and urban 600, size 8
9	three clusters, rural, mixed and urban 600, size 16
10	Global clustering, twin 600, fixed distance 0
11	Global clustering, twin 600, fixed distance 0.01
12	Global clustering, twin 600, exponential distance 0.04
13	Global clustering, triple 600, fixed distance 0
14	Global clustering, triple 600, fixed distance 0.02
15	Global clustering, triple 600,exponential distance 0.08

Figure 4.9: Power comparison between Kulldorff's and the Semiparametric with likelihood ratio test methods for binary type data using the northeastern US benchmark data with 600 simulated cases.



No.	Kull.paper	Kull me	SemiwFDR=0.1	SemiwFDR=0.05	Bonferroni
0	0.050	0.050	0.054	0.027	0.004
1	0.885	0.885	0.843	0.794	0.711
2	0.920	0.913	0.870	0.842	0.793
3	0.923	0.917	0.965	0.930	0.761
4	0.968	0.970	0.941	0.897	0.857
5	0.981	0.978	0.960	0.941	0.907
6	0.991	0.992	0.996	0.987	0.935
7	0.999	1.000	0.998	0.997	0.996
8	0.991	0.992	0.989	0.978	0.913
9	0.980	0.979	0.983	0.965	0.837
10	0.826	0.821	0.678	0.563	0.399
11	0.277	0.268	0.335	0.229	0.072
12	0.180	0.192	0.161	0.161	0.047
13	0.966	0.994	0.984	0.954	0.903
14	0.313	0.326	0.425	0.351	0.138
15	0.166	0.174	0.266	0.175	0.045

No.	Scenario
0	No cluster 6000, under null hypothesis (type I error)
1	single cluster, mixed 6000, size 1
2	single cluster, rural 6000, size 4
3	single cluster, urban 6000, size 16
4	two clusters, mixed and urban 6000, size 1
5	two clusters, rural and urban 6000, size 4
6	two clusters, rural and mixed 6000, size 16
7	three clusters, rural, mixed and urban 6000, size 1
8	three clusters, rural, mixed and urban 6000, size 8
9	three clusters, rural, mixed and urban 6000, size 16
10	Global clustering, twin 6000, fixed distance 0
11	Global clustering, twin 6000, fixed distance 0.01
12	Global clustering, twin 6000, exponential distance 0.04
13	Global clustering, triple 6000, fixed distance 0
14	Global clustering, triple 6000, fixed distance 0.02
15	Global clustering, triple 6000, exponential distance 0.08

Figure 4.10: Power comparison between Kulldorff's and the Semiparametric with likelihood ratio test methods for binary type data using the northeastern US benchmark data with 6000 simulated cases.

as containing more cluster regions, the power of both methods increases, which demonstrates the validity of the two methods. For instance, scenario No. 6 in figure 4.9 has a total of 600 cases and two cluster regions where each cluster region contains 16 counties. The power of Kulldorff's method is 0.996, whereas the power of semiparametric method with q-value significance level of 0.1 and 0.05 obtain the power 0.996 and 0.993, respectively, which is quite close to the power of Kulldorff's. The semiparametric method with Bonferroni correction is 0.968, which is expected to be the lowest, but still is quite reasonable.

For scenarios No. 10 to 15, both figures show that the two methods do not do very well compared with the results in hot-spot detection. This is so because both Kulldorff's and the semiparametric methods are not designed to detect global clustering pattern. The figures show that Kulldorff's method is slightly more powerful than semiparametric method with an exception of scenario No. 14. It is also shown that for both methods, the larger the r is, the lower is the detection power.

4.4 Comprehensive Power Study: Non-binary Data

4.4.1 Non-binary Data Set and Simulation Plan

For non-binary data scan, we use the simulated ordinal categorical data with one data point corresponding to one observation. The data are aggregated to state levels distributed in 18 states. Most of them are middle south states, including Alabama (AL), Arkansas (AR), Texas (TX), Virginia (VA), etc. Each state is

graphically represented by its centroid coordinate. Figure 4.11 shows the map of the states included in our simulated study.

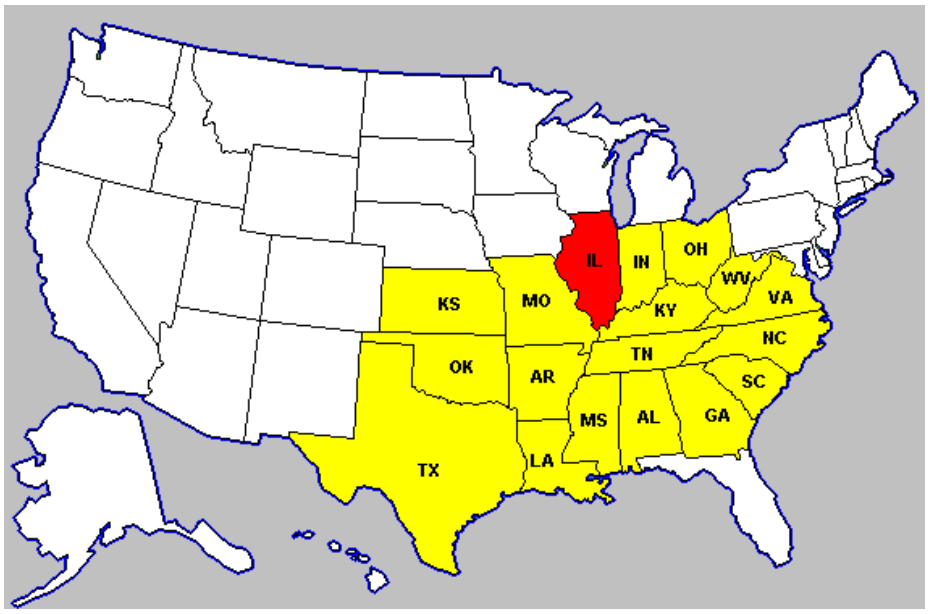


Figure 4.11: Map showing the states included in the simulation denoted with color and the abbreviation of state names. The state Illinois with red color is illustrated as one possible cluster region in our simulated data.

The ordinal categorical data are integer data generated from quantized normal data obtained from the integer part of the original normal data. We refer to quantized normal II and quantized normal III data as described in the limited power study section. Also see Kedem and Wen (2007) [26]. The quantized normal II data are derived from normal data with the same mean but different variance inside and outside the cluster region, whereas the quantized normal III data are derived from normal data with both different mean and variance inside and outside the cluster region. To see how Kulldorff's and the semiparametric methods perform when the

difference between the cluster and the non-cluster region is small, namely a more difficult cluster detection problem, the data set "Quantized Normal III (small)" was generated. Table 4.1 lists the mean and variance parameters used to generate the quantized normal data.

Table 4.1: Parameters for Simulating the Ordinal Categorical Data from Quantized Normal

Data Type	***Inside the cluster***	***Outside the cluster***
Quantized Normal II	$\mu = 13$, $\sigma^2 = 8$	$\mu = 13$, $\sigma^2 = 4$
Quantized Normal III	$\mu = 7.2$, $\sigma^2 = 13$	$\mu = 6$, $\sigma^2 = 9$
Quantized Normal III (small)	$\mu = 6.5$, $\sigma^2 = 13$	$\mu = 6$, $\sigma^2 = 9$

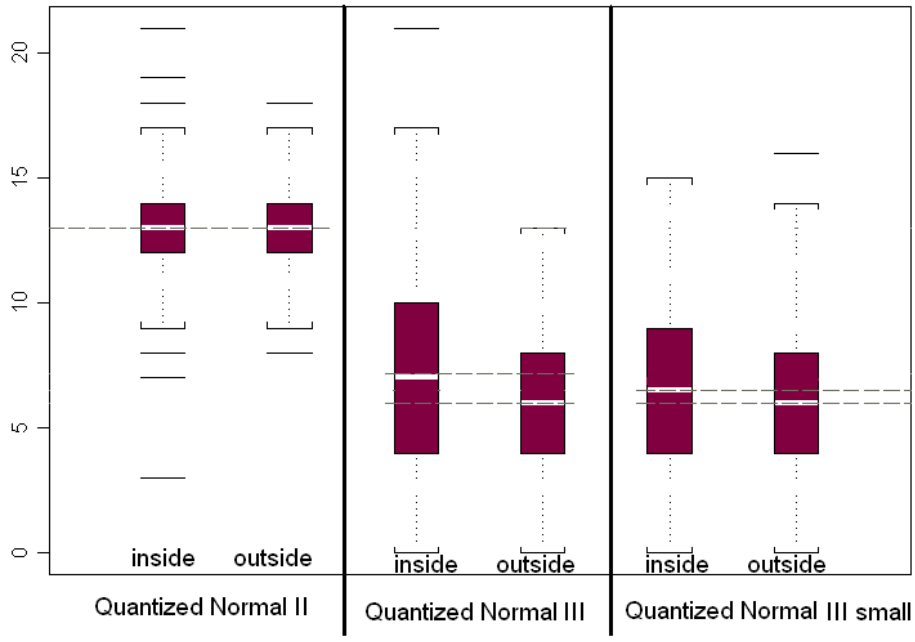


Figure 4.12: Box Plots of the Simulated the Ordinal Categorical Data

For each type of data, the average sample size within each state is around 130, hence there is a total of $130 \times 18 = 2340$ observations in each generated data set. To perform power comparison within a reasonable time scale, we generated 100 random data sets with a single cluster and also multiple clusters for both Quantized Normal II and Quantized Normal III data, respectively. The single clusters means only one state was randomly chosen as the cluster region. By multiple clusters we mean four states were randomly chosen as the cluster region, where the four states are not necessarily contiguous. Notice that multiple clusters constitute a stronger clustering pattern which in general is easier to detect.

In our power comparison using these non-binary ordinal categorical data, the same data sets were used for both methods. Kulldorff's method was applied with the Poisson model (inappropriate model) and with the ordinal model (correct model), while the semiparametric method was applied with the vector tilt function $h(x) = (x, x^2)$. We used Kulldorff's *SaTScan* v7.0.1 software to conduct the cluster detection by the ordinal model. The results of the power comparison for the ordinal categorical data are shown is Figures 4.13 and 4.14. Observe that it is appropriate to choose $h(x) = (x, x^2)$ for binary data as well, because in that case the coefficient of x^2 term is 0. See more details in the discussion section.

4.4.2 Results for Ordinal Categorical Data

The results of power comparison for non-binary ordinal categorical data using the generated middle south US data set are shown in Figures 4.13, 4.14 and 4.15.

All the tests are two-sided tests. The significance level for Kulldorff's method is still 0.05, and the q-value significance level for the tests in semiparametric method is also set at 0.05. This means the true type I error level of the semiparametric method is lower than Kulldorff's as explained above.

For ordinal categorical data which are generated from the quantized normal II type, figure 4.13 shows that semiparametric method with the likelihood ratio test has the highest power of detecting potential clusters among all the tests. The semiparametric method with the χ_1 test works well but not as well as the likelihood ratio test. This is in line with the limited power study in Kedem and Wen (2007) [26] who showed the likelihood ratio test was the most powerful tests among three tests from the semiparametric density ratio model. Moreover, when the clustering pattern is stronger, in this case, when the study region contains multiple clusters, the power of χ_1 test increase to 0.94, which is almost close to the power of the likelihood ratio test. The power of Kulldorff's method with Poisson model is very low, because the Poisson model is inappropriate when the variance is not constant Kulldorff's method with ordinal model is comparable to the semiparametric method with the χ_1 test in detecting potential clusters.

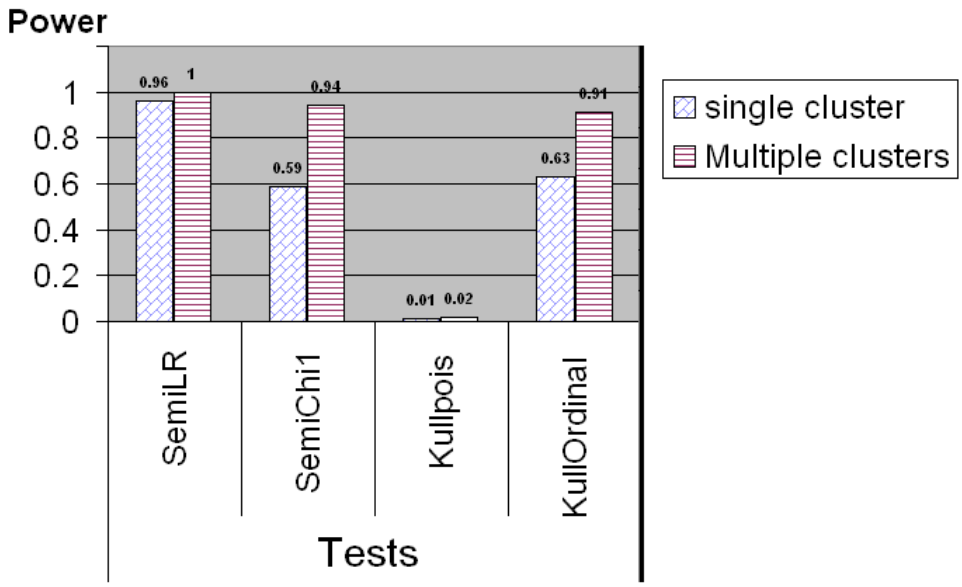


Figure 4.13: Power comparison between Kulldorff's and the Semiparametric methods for ordinal categorical data generated from quantized normal II data, where the means are the same but the variances are different, between the cluster region and the rest of the area.

For ordinal categorical data which are generated from the quantized normal III type, figure 4.14 shows that semiparametric method performs in the same way as in the quantized normal II case. The likelihood ratio test still has the highest power in both single and multiple clusters situation. The power of all tests increases as the clustering pattern becomes stronger. Kulldorff's method with the ordinal model works as well as the semiparametric method. Kulldorff's method with the Poisson model works but less powerful as compared with the other tests. This is because for the quantized normal III data, Kulldorff's Poisson model can detect changes in the mean while ignoring changes in variances.

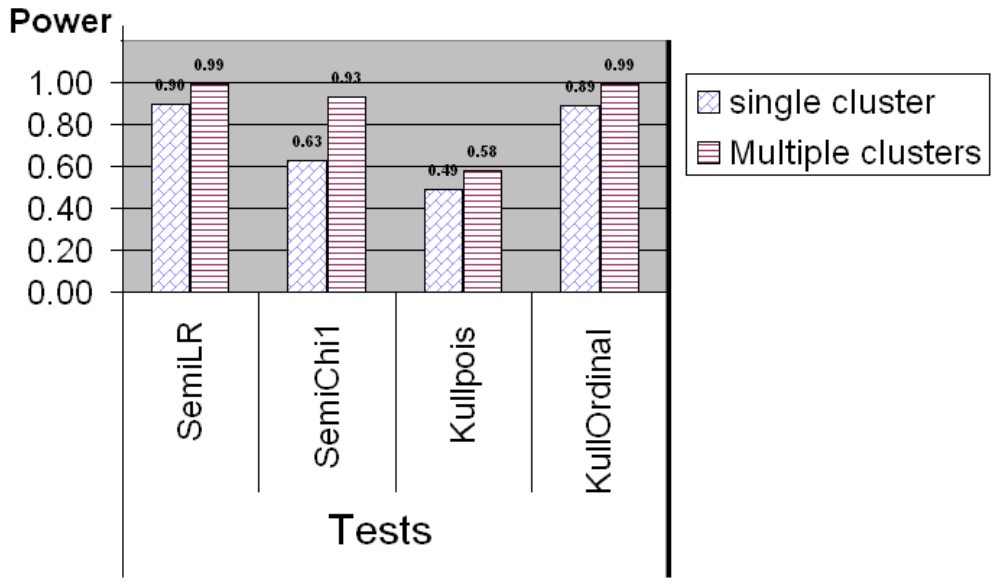


Figure 4.14: Power comparison between Kulldorff's and the Semiparametric methods for ordinal categorical data generated from quantized normal III data, where both means and variances are different inside and outside the cluster region.

Figure 4.15(a) shows the power results when the difference between the cluster and non-cluster regions is small. The data are still ordinal categorical data generated from the quantized normal III data. However, this time the difference of the mean in the cluster region is set to be very close to the the non-cluster region, while the variance inside and outside the cluster region is kept the same as in the previous case (refer to Table 4.1 for the simulation parameters). In this way, the cluster is more difficult to detect. Not surprisingly, for the single cluster case, the detection power of both Kulldorff's and the semiparametric method significantly decreases due to the weaker cluster pattern. The power for the likelihood ratio test from the semiparametric method and Kulldorff's method with the ordinal model continue achieving the highest power, although it is much lower than in the previous case where the differences inside and outside the cluster region are substantial. The power of the χ_1 test of the semiparametric method also decreases, and Kulldorff's method with the Poisson model continues yielding the lowest power which is almost 0. The multiple cluster case is similar to the single cluster case but with a higher power.

Interestingly, if we look at the accuracy, which means detecting the true cluster region correctly with its exact size, it is shown as in figure 4.15(b) that the most accurate method is the semiparametric method with the likelihood ratio test statistic, which has a more than 50% higher accuracy rate than Kulldorff's method with the ordinal model.

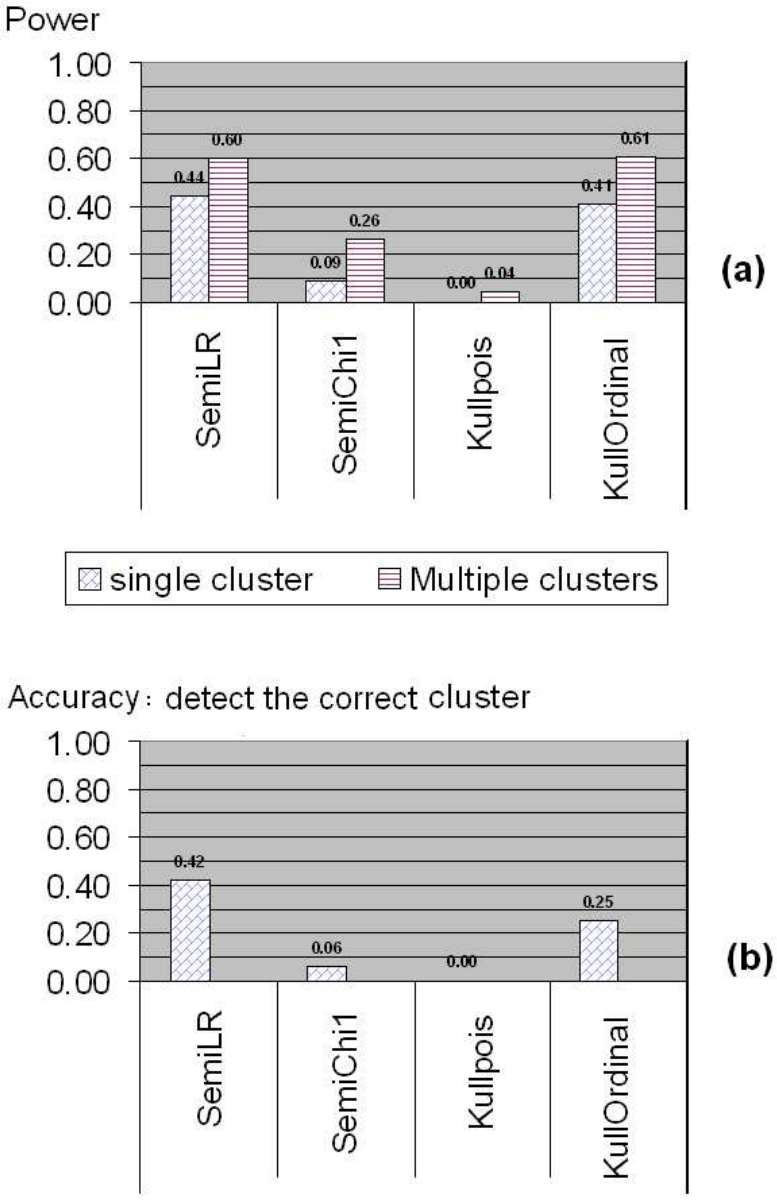


Figure 4.15: Power comparison between Kulldorff's and the Semiparametric methods for ordinal categorical data with small differences. The data are in quantized normal III small type. (a) Existence. (b) Accuracy.

Chapter 5

Data Analysis examples

Chapter 3 illustrates the Semiparametric cluster detection method using simulated data. In Chapter 4, it is shown that the Semiparametric method achieves comparable power to that of the celebrated Kulldorff's method. In some cases, the Semiparametric method has a higher power. In this chapter, I apply the Semiparametric cluster detection method to real data.

5.1 North Humberside Childhood Leukemia Data

Both the Semiparametric method with the likelihood ratio test using tilt function $h(x) = x$ and Kulldorff's method with the Bernoulli model are applied to a real data set from Kulldorff's *satscan* website `http://www.satscan.org/`. It gives the spatial location of 62 cases of childhood leukemia and lymphoma in North Humberside, England, between 1974 and 1986, as well as 141 controls (Alexander et al., 1990) [1]. The scientific question is to see if there is some region with a higher disease rate. A snapshot of part of the data set is shown in Figure 5.1, and the spatial locations of the region's postal zones are shown in Figure 5.2(a).

A circular scanning window was used and moved across all postal zones with a variable size ranging from roughly the size of a postal zone to no more than 20% of the study region. Both Kulldorff's scan statistics and the Semiparametric density

Postal	Xcoord	Ycoord	blpop	blcases
133	5123.00	4303.00	1.00	0.00
134	5123.00	4315.00	1.00	0.00
135	5124.00	4317.00	1.00	1.00
136	5124.00	4332.00	1.00	0.00
137	5129.00	4309.00	1.00	0.00
138	5129.00	4326.00	1.00	0.00
139	5130.00	4310.00	1.00	0.00
140	5131.00	4305.00	1.00	1.00
141	5131.00	4315.00	1.00	0.00
142	5132.00	4307.00	1.00	1.00
143	5134.00	4319.00	1.00	1.00
144	5135.00	4303.00	1.00	0.00
145	5136.00	4303.00	1.00	0.00
146	5138.00	4327.00	2.00	1.00
147	5139.00	4308.00	1.00	0.00
148	5140.00	4305.00	1.00	1.00
149	5140.00	4326.00	2.00	1.00
150	5141.00	4332.00	1.00	0.00
151	5147.00	4654.00	1.00	1.00
152	5149.00	4298.00	1.00	1.00
153	5150.00	4303.00	1.00	0.00
154	5150.00	4308.00	1.00	1.00
155	5150.00	4310.00	1.00	1.00

Figure 5.1: Snapshot of part of the North Humberside childhood leukemia and lymphoma data set

ratio method point to the same cluster shown in Figure 5.2(b), consisting of postal zones (14, 18, 19, 26), as the primary cluster candidate. However, from the software SatScan version 7.0.1 described at `http://www.satscan.org/`, Kulldorff's method gives a p-value of 0.674, nonsignificant, whereas the Semiparametric p-value without adjusting for multiple testing is 0.002. The Semiparametric method coupled with the FDR control gives a q-value of 0.073, which is on the boundary, suggesting that the detected cluster could be a true cluster. Thus, the two approaches lead to very different conclusions as expressed by very different p-values. Which one is correct?

Working with the same data, Cuzick et al. (1990) found that the true cluster is likely to consist of 4 postal zones [1]. Moreover, environmental studies by Colt and Blair (1998) and Mckinney et. al. (1991) reported that the association between childhood leukemia and paternal exposure to solvents was quite strong, and that a global cluster was located in North Humberside [5, 41]. Thus, the results from the Semiparametric method are more in line with the medical and environmental studies, and the located cluster candidate as well as its significance given by the Semiparametric method seems more credible. More conclusive results may be obtained by increasing the sample size.

5.2 Maryland-DC-Virginia Crime Data

Besides cancer research and epidemiology studies, another important application area of cluster detection is in crime mapping [12, 37]. Actually criminology has a long history of using mapping techniques, such as "colored pin maps", to help

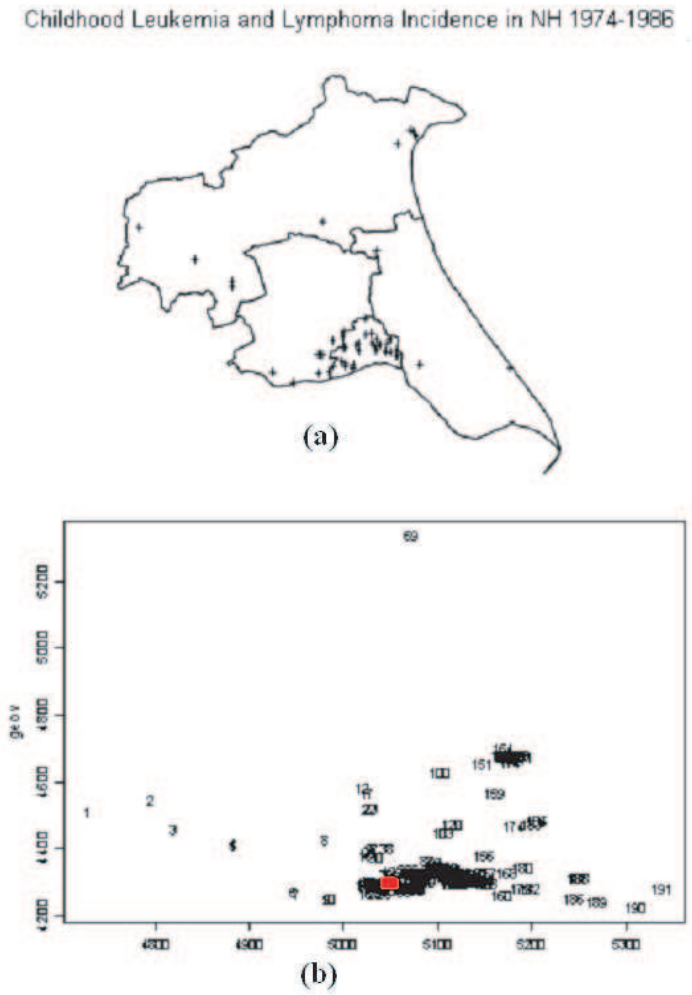


Figure 5.2: (a.) The geographical map. (b). The detected cluster candidate in red.

police officers improve public safety. For instance, through mapping crime occurrences, police officers or investigators can determine regions with high crime rate, or figure out the route of drug flow, and so on. This helps the criminal justice or law enforcement specialists to optimize the allocation of resources. In recently years, the advance of computer technology as well as geographic information systems (GIS) make crime mapping widely available [55, 39, 69].

In this section, we apply our Semiparametric cluster detection method with both the χ_1 test and likelihood ratio test to detect crime clusters (both hot-spot and cold-spot), if any, using the the 2001 - 2004 data set of the annual number of arrests in Maryland-DC-Virginia since it is believed that high arrest rates indicates high crime rates. The data are from National Consortium on Violence Research (NCOVR) website, which is a research, training, and data resource specializing in violence research. NCOVR was funded in 1995 by a grant from the National Science Foundation in cooperation with the National Institute of Justice. For more information of NCOVR, see its website at (`http://www.ncovr.heinz.cmu.edu/`).

The annual number of arrests data I used in this case study are aggregated to county level (159 counties in total). To adjust the population, I used the 2002 Population data and assume it is fixed throughout the period 2001 to 2004. The spatial coordinates are the longitude and latitude of the center of each county. In reality, there were minor changes in the population during four years, but the changes were small relative to the total population, and it doesn't hurt to regard the population as constant in a short period. Figure 5.3 gives a snapshot of the crime

data in year 2001.

SiteNum	Counties	Arrest2001	Pop2002	Long	LAT	Rate2002	Period
1	Allegany	5798.00	77134.00	-78.69	39.60	0.08	2001
2	Anne Arundel	26819.00	504601.00	-76.59	39.16	0.05	2001
3	Baltimore	35260.00	141498.00	-76.64	39.47	0.25	2001
4	Baltimore (City)	95025.00	635815.00	-76.60	39.36	0.15	2001
5	Calvert	5449.00	76838.00	-76.56	38.54	0.07	2001
6	Caroline	2195.00	30681.00	-75.83	38.87	0.07	2001
7	Carroll	4356.00	155503.00	-77.02	39.56	0.03	2001
8	Cecil	4969.00	88574.00	-75.95	39.59	0.06	2001
9	Charles	7808.00	124225.00	-77.00	38.56	0.06	2001
10	Dorchester	2384.00	31610.00	-76.04	38.56	0.08	2001
11	Frederick	9152.00	201237.00	-77.40	39.48	0.05	2001
12	Garrett	1705.00	30757.00	-79.29	39.53	0.06	2001
13	Harford	7891.00	225262.00	-76.31	39.56	0.04	2001
14	Howard	9650.00	255406.00	-76.95	39.24	0.04	2001
15	Kent	1325.00	19783.00	-76.04	39.26	0.07	2001
16	Montgomery	17672.00	899994.00	-77.18	39.18	0.02	2001
17	Prince George's	26998.00	823314.00	-76.84	39.01	0.03	2001
18	Queen Anne's	1773.00	41801.00	-76.03	39.05	0.04	2001
19	Somerset	1313.00	25503.00	-76.03	38.11	0.05	2001
20	St Mary's	5670.00	88842.00	-76.59	38.29	0.06	2001
21	Talbot	2020.00	34844.00	-76.10	38.77	0.06	2001
22	Washington	7505.00	135950.00	-77.82	39.65	0.06	2001

Figure 5.3: Snapshot of part of the Maryland-DC-Virginia Crime data set

A circular window is used to scan the whole region aiming to detect both high and low risk clusters. The window size is varied ranging from roughly the size of including one county to a maximum of 50% of the study region. The results from the Semiparametric cluster detection method are listed in Table 5.1. It shows that from year 2001 to 2004, although the arrest rate (number of arrests divided by the population) is changing, the primary high crime risk cluster always includes Baltimore county and Baltimore city centered at Baltimore county. The average arrest rate in this Baltimore cluster is four to five times higher than the rest of region. The p-value and q-value are both less than 0.001 showing significant high

crime risk. The primary low risk cluster includes Montgomery county, Howard county, Fairfax county, District of Columbia, Falls Church (City), Arlington county, Fairfax City, Alexandria City, Prince George's county, Loudoun county, centered at Montgomery county. The p-value and q-value are also both less than 0.001 showing significant low crime risk. Figure 5.4 shows the primary high and low crime risk clusters.

Table 5.1: Results of High and Low Crime Risk Cluster from Yr 2001∼2004

Primary Cluster	**High Risk**	**Low Risk**
Cluster center	Baltimore County	Montgomery county
Cluster size	2 counties and cities	$12 \sim 15$ counties and cities
Relative arrest rate Ratio	$4.31 \sim 5.05$	$0.31 \sim 0.36$
p-value (from χ_1 and LR)	< 0.001	< 0.001
q-value (from χ_1 and LR)	< 0.001	< 0.001

Figures 5.5 to 5.8 show the arrest rates of each county for each year. The figures also demonstrate the constancy of the primary high and low risk clusters.

The results are not surprising. They are all consistent with the economic and demographic factors. Throughout 2001 to 2005, on average, the arrest rate of Baltimore is four to five times higher than the average of the rest of the three states. Moreover, this ratio appears to be increasing. According to crime statistics there were 269 homicides in Baltimore in 2005, giving it the highest homicide rate per 100,000 of all U.S. cities of 250,000 or bigger population [7]. The homicide

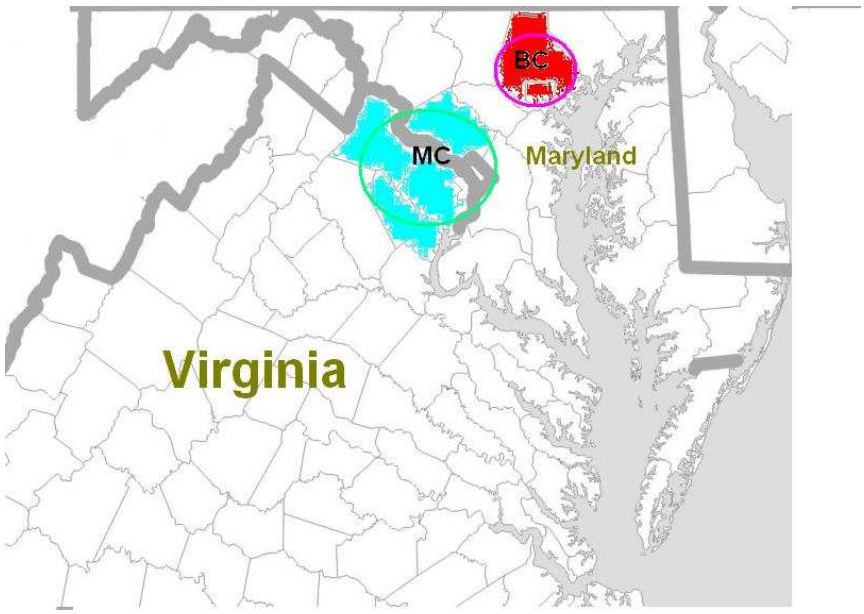


Figure 5.4: Maryland-DC-Virginia High and Low Risk Crime Cluster. Red means high crime risk cluster, Navy means low crime risk cluster.

rate in Baltimore is nearly seven times the national rate, six times the rate of New York City, and three times the rate of Los Angeles. In 2007, the CNN/Morgan Quitno "Most Dangerous City" Rankings (2007) ranks Baltimore as the 12th most dangerous American city. Baltimore is second only to Detroit among cities with a population over 500,000 [8, 9]. The high risk of crime has troubled Baltimore for years. The main reasons for the high crime rate in the Baltimore area are illegal drug trade, dreadful public schools, and lack of jobs. Besides strengthening the police force, it is essential to improve the education level of the school system, cut off drug trade lines, and provide more jobs.

On the other hand, the Montgomery County cluster as well as its nearby counties including Fairfax County, etc. have a national reputation for their public educa-

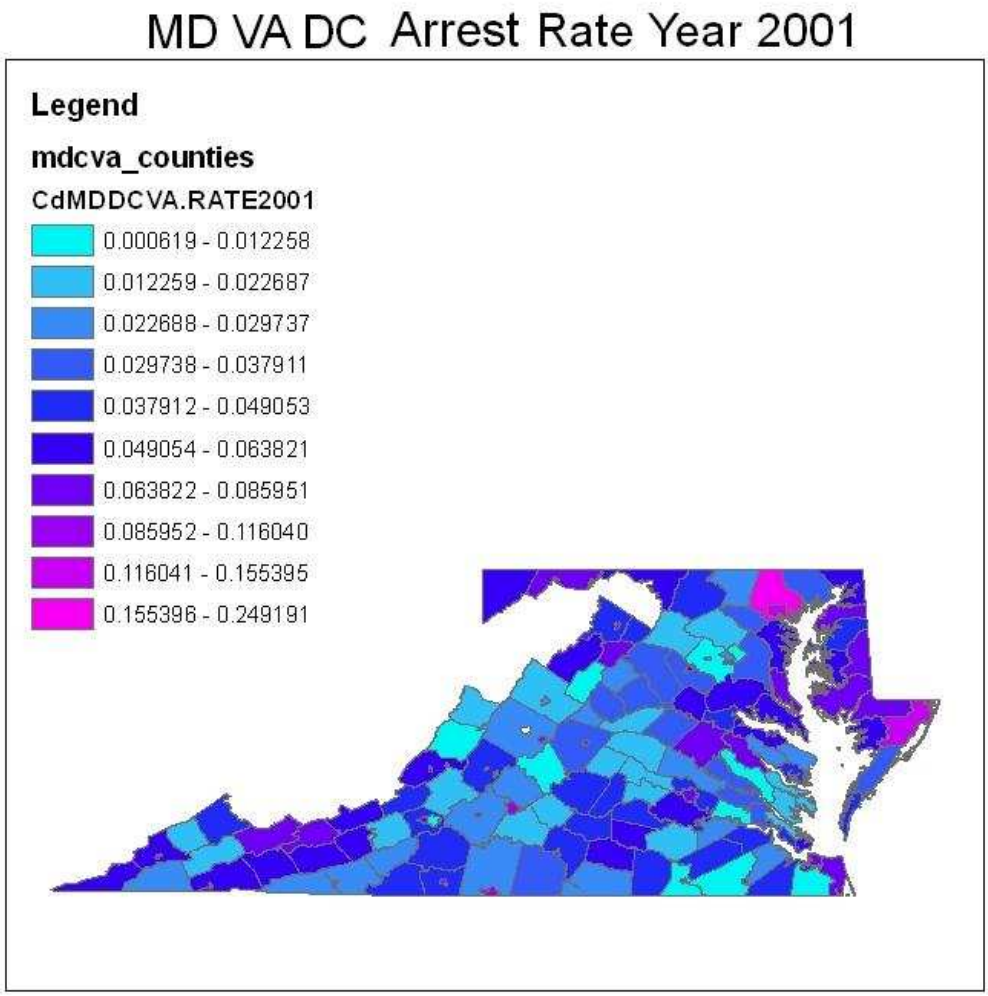


Figure 5.5: Maryland-DC-Virginia Arrest Rate by County in Year 2001

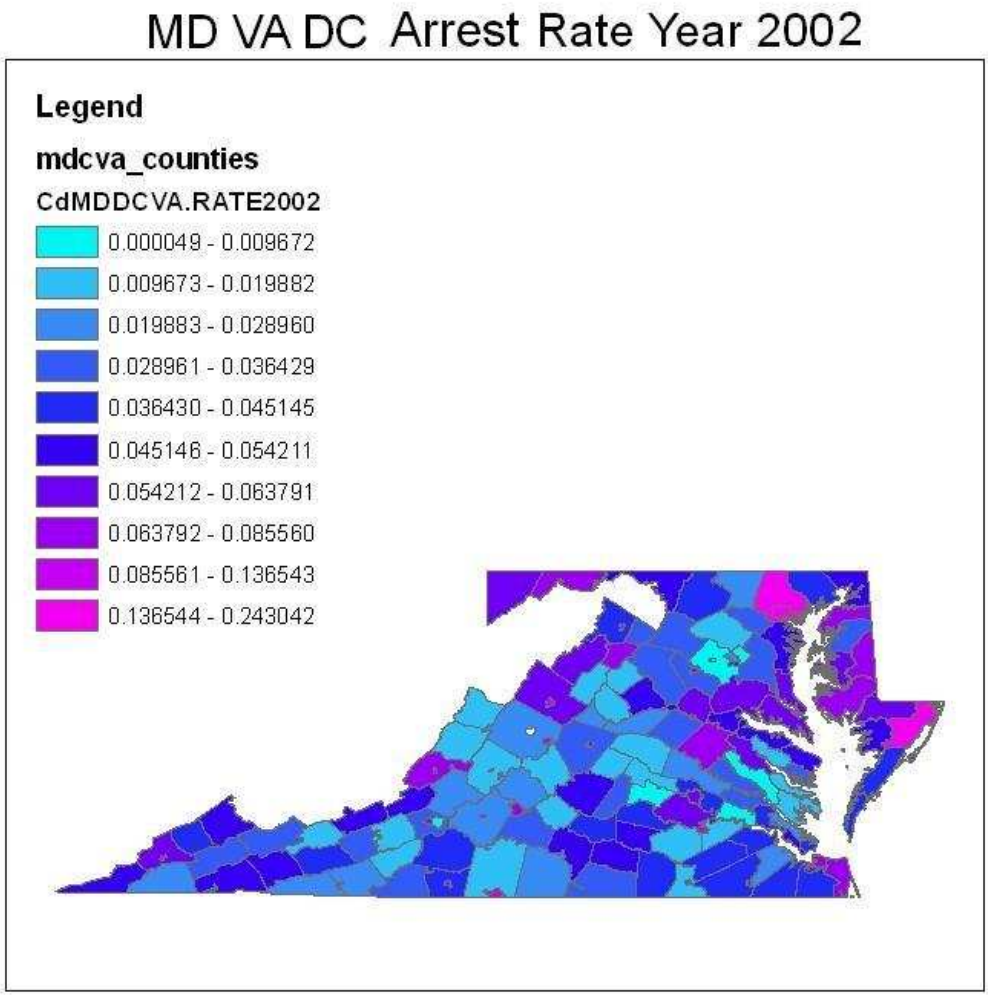


Figure 5.6: Maryland-DC-Virginia Arrest Rate by County in Year 2002

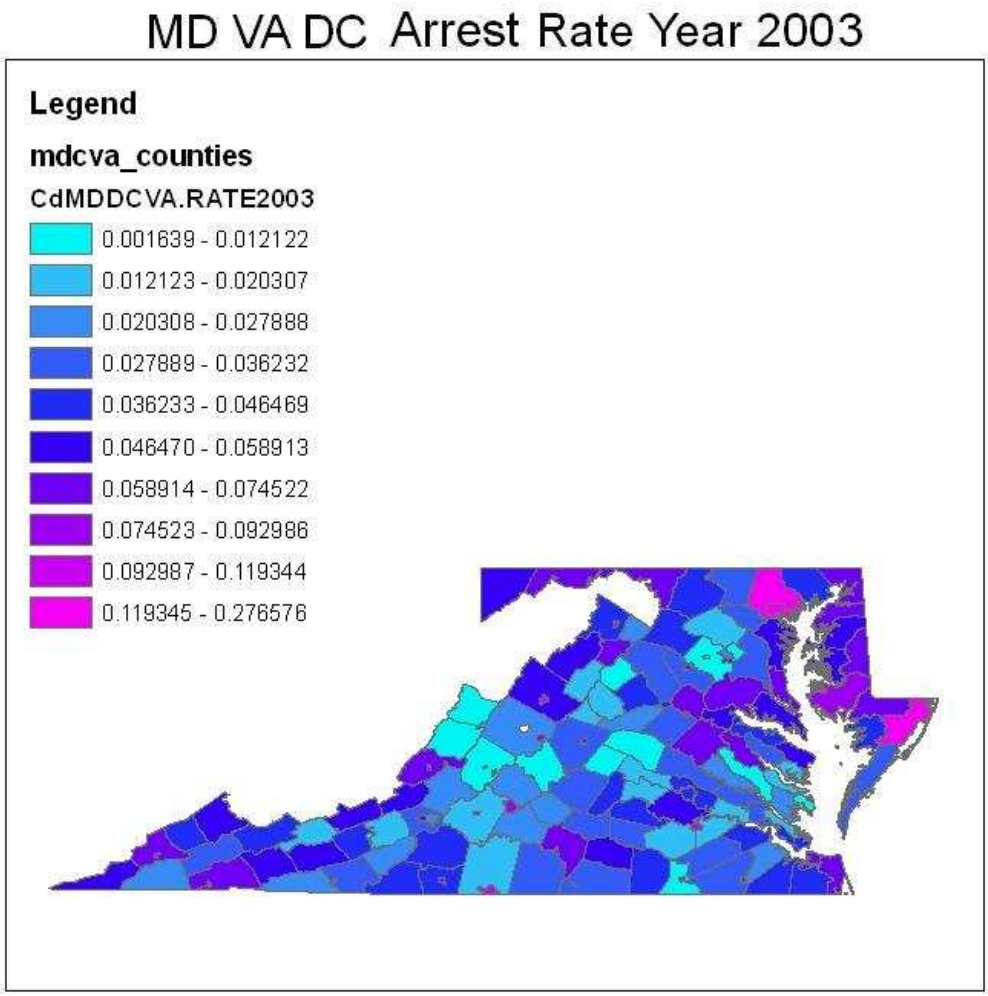


Figure 5.7: Maryland-DC-Virginia Arrest Rate by County in Year 2003

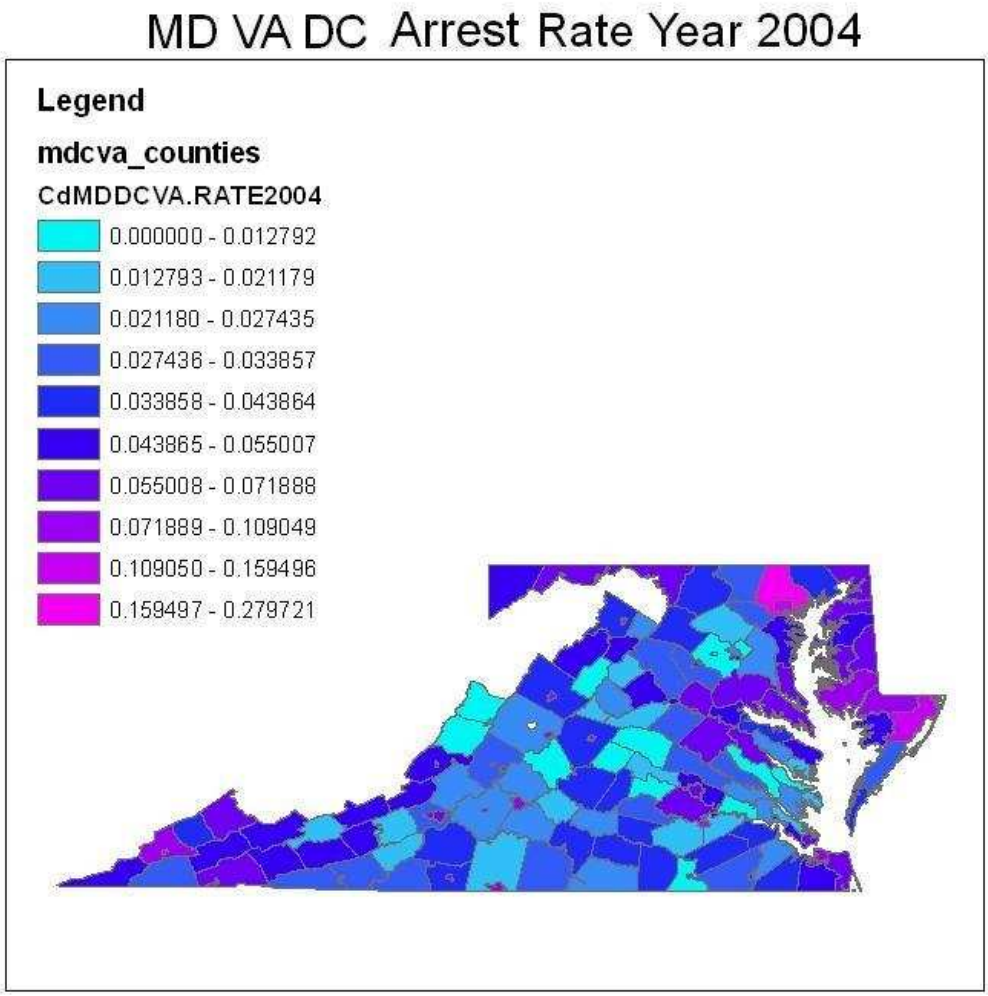


Figure 5.8: Maryland-DC-Virginia ArrestRate by County in Year 2004

tion system. Students in Montgomery county public schools score among the top in the United States on Advanced Placement Examinations [71]. The Fairfax County government spends more than half of its fiscal budget on its education system. Its county public school system contains the Thomas Jefferson High School for Science and Technology (TJHSST), a Virginia Governor's School. TJHSST consistently ranks at or near the top of all United States high schools due to the extraordinary number of National Merit Semi-Finalists and Finalists, the high average SAT scores of its students, and the number of students who annually perform nationally recognized research in the sciences and engineering [72]. In addition, Montgomery County has a very lucrative business climate. It is the epicenter for biotechnology in the Mid-Atlantic region, the third largest biotechnology cluster in the nation. There are many large firms and federal agencies located in Montgomery County, including Lockheed Martin, Marriott International, BAE Systems Inc, Genentech, National Institutes of Health (NIH), Food and Drug Administration (FDA), National Institute of Standards and Technology (NIST), and so on. Those companies and organizations attract a lot of highly educated work force and provide tremendous work opportunities. Thus, it makes the crime rate here remains consistently low.

Interestingly, DC and Prince George's county are also included in this low crime risk cluster. It seems a little different than what you expect because these two regions are known as violent areas. But the data show that the arrest rates in DC and PG county are 0.9% and 3%, respectively. Both of them are less than 4.7%,

the overall arrest rate of the three states. This may be due to a large population, although the absolute number of arrests are high. Another reason could be that it is hard for the circular scanning window also to delineate the exact boundary of the cluster.

It is also noticed that Worcester County of Maryland where Ocean City is located is a region with a second highest arrest rate, which is almost three times than the overall average. However, the results from the Semiparametric cluster detection method doesn't show that Worcester County is a secondary cluster candidate. The high arrest rate in Worcester County may due to its small population.

Chapter 6

Summary and Discussion

In this dissertation, I develop a cluster detection approach by using a Semiparametric method applied to moving windows of variable size as suggested by Kulldorff's method. The only assumption needed regards the exponential tilt function $\boldsymbol{h}(x)$, but unlike Kulldorff's method, no specific distributional assumptions are necessary, and the testing procedure with $m = 2$ is quite simple. Likewise, there is no need to know the number of cases a priori. The practical potential of the method was demonstrated with real and artificial data. It successfully detects potentially high risk clusters (hot-spot) as well as low risk clusters (cold-spot). In addition, the significant tests of the Semiparametric method use χ_2 test to obtain the p-value directly, so there is no need to run the time consuming Monte Carlo testing procedure. As an example, for the Maryland-DC-Virginia crime data set in Chapter 5, using my *acer* laptop, Intel *Centrino* 1.6GHz CPU, 512M memory, the Semiparametric method in the Splus environment costed around 8 seconds to get the final results. But Kulldorff's SatScan v7.0.1 software costed around 50 seconds to get the results.

The results of the power study show that when detecting localized clusters, both the Semiparametric and Kulldorff's method achieve comparably good power. For binary population-case data, Kulldorff's method with the Poisson model may have a slightly higher power than the Semiparametric method with tilt function

$h(x) = x$. For non-binary data, such as ordinal categorical data, the Semiparametric method with tilt function $\boldsymbol{h}(x) = (x, x^2)'$ is slightly more powerful than Kulldorff's method with an ordinal model. If Kulldorff's method is applied with an inappropriate model, that is using the Poisson model to analyze ordinal categorical data, it may fail to detect any potential clusters. For instance, Kulldorff's method with the Poisson model obtains a very low power for quantized normal II data, while it still works for quantized normal III data, but with a relatively lower power compared with the ordinal model. We also find that in our Semiparametric method the likelihood ratio test seems to have a higher power than the χ_1 test in detecting potential clusters. When the localized clustering pattern is strong, for instance, multiple cluster regions or the difference inside and outside the cluster region is large, both tests obtain good power. When the clustering pattern is weak, that is, the difference is not that large, the likelihood ratio test seems to be more acuminous, while the χ_1 test could be insensitive to undesired fluctuations. On the other hand, the Likelihood ratio test only can test "either high or low values", but χ_1 test can be easily transformed into a one-sided test if the tilt function has a scalar form. In practice, it is prudent to use both tests whenever possible for potential clusters.

In the power study for non-binary data, we only used the ordinal categorical data, which are integer data, to compare the power of the two methods. However, Kulldorff's method also offers the exponential model to analyze survival time data and a normal model to analyze continuous data. A future study could compare the power of the Semiparametric method and Kulldorff's method for continuous data.

We expect the Semiparametric method to work well, since the Semiparametric density ratio model was originated designed for continuous data. In addition, Semiparametric model provides a more consistent setup than Kulldorff's method. Kullorff's method requires different models, namely different scan statistics, for different types of data, whereas the Semiparametric method requires no specific distributional assumptions except for the exponential tilt function $\boldsymbol{h}(x)$. In practice, the choice of $\boldsymbol{h}(x) = (x, x^2)'$ is appropriate for many types of continuous and discrete data.

If the underlying distribution is known exactly, we choose the true tilt function $\boldsymbol{h}(x)$ to get the best performance. For instance, if the data are from Bernoulli or Poisson distribution, we use the tilt function $h(x) = x$; if the data are from normal, we use $\boldsymbol{h}(x) = (x, x^2)'$; if the data are from Gamma distribution, we can use $\boldsymbol{h}(x) = (x, \log x)'$. We may choose the tilt function as $\boldsymbol{h}(x) = (x, x^2)'$ for binary data as well although the x^2 term is not necessary. Appendix A.2 shows that the power of the Semiparametric method does not change since the parameter associated with x^2 is 0. Simulation studies have demonstrated that if a term in the tilt function is not necessary, the parameter associated with that term is close to 0 also. A clue of how to choose a satisfactory $\boldsymbol{h}(x)$ for a given situation can be derived from common exponential families (recall Chapter 3). If the underlying distribution is not known, there could be a problem of a misspecified tilt function. Fokianos and Kaimi (2006) demonstrates that a misspecified $\boldsymbol{h}(x)$ could decrease the power of the corresponding tests [14, 15]. Yet, there are examples where very different choices of $\boldsymbol{h}$ could lead to similar test results. For instance, in an application to meteorological data in Kedem

et al. (2004), the choice of $h(x)$ as x or $\log x$ led to very similar test results [25]. To check whether the assumption of the exponential tilt density ratio model holds, Qin and Zhang (1997) [53] propose a Kolmogorov-Smirnov type statistic to test the goodness of fit of the density ratio model for two-sample case, which also gives a guidance of an appropriate choice of $\boldsymbol{h}(x)$. Lu (2007) extend this goodness of fit test to m-sample case [40].

Interestingly, Kulldorff's method and the Lawson-Waller focused score test may still perform well even for non-Poisson count data as long as the variance of the observations over the study region does not change much. However, these two methods seem to lose power when the variance changes appreciably over the region. As for the Semiparametric method, it seems that for a non-homogeneous regional variance the choice of $\boldsymbol{h}(x) = (x, x^2)'$ suggested by the normal distribution is sensible. Shmueli et al. (2006) [57] revive the Conway-Maxwell-Poisson (CMP) distribution to fit discrete data. The CMP distribution is a two-parameter extension of the Poisson distribution that generalizes some well-known discrete distributions (Poisson, Bernoulli and geometric). In this sense, $h(x) = (x, logx!)$ derived from the CMP distribution could also be used as an alternative to $h(x) = (x, x^2)$ for discrete data.

The Semiparametric density ratio model essentially tests the homogeneity or equidistribution of two or more samples, therefore, besides Kulldorff's circular scan window, the Semiparametric method may also adapt to other shapes of the scanning window or scanning schemes, such as the elliptic window scan, Patil and Taillie's up-

per level set scan and Tango's flexible scan mentioned in Chapter 1. More precisely, in scanning for clusters, and regardless of regular or irregular shapes of the scanning window, as long as the window separates the whole study region into two samples, one inside the window and one outside the window, the Semiparametric method can be applied. However, the Semiparametric method ignores the information about the location of a cases except whether a case is inside or outside the current window. Thus the Semiparametric method may not have good power for global type clustering clusters as shown in scenarios 10 to 15 in Figure 4.9 and Figure 4.10 where clustering occurs throughout the study region.

It is also important to keep in mind that whatever the shape of the most likely cluster, it only indicates the general area of the true underlying cluster, and that the exact boundary of the detected clusters is uncertain. This is sufficient for most practical purposes, as the Semiparametric cluster detection method's main purpose is to generate a signal with a general idea of where an outbreak or higher than normal activity has occurred. More detailed information about the outbreak, its cause, nature and extent, can only be obtained through detailed investigations by specialists in corresponding areas, who should not only focus on the area within the most likely cluster, but also on neighboring localities. The exact choice of shapes is not of critical importance. If computing resources allow, better results may be obtained using an irregular rather than a circular scan window, depending on the shape of the true underlying cluster. For valid statistical inference, it is important that the choice of the scan window is made a priori though, before analyzing the

data, in order to avoid pre-selection bias.

A limiting factor of the complete power study in this dissertation is that for non-binary type data we only used 100 runs at one point for each power comparison. That is because it is tedious and time consuming to run SatScan software and document the results manually. In addition, the Semiparametric method also takes longer time to numerically estimate the parameters α and $\boldsymbol{\beta}$ especially when the combined sample size is large, while for binary scan the MLEs of α and β are available in a closed form. However, although 100 doesn't sound like a large number in simulation, the results are reasonable. From Figures 4.13 and 4.14, it is already clear that the Semiparametric method achieves comparable good power as Kulldorff's method with the correct model. If Kulldorff's method is applied with an inappropriate model, the power may decrease a lot. Figure 4.15 demonstrates that when the difference between the cluster and non-cluster region is small, the detecting power for both methods decreases, but the likelihood ratio test of the Semiparametric method seems better in term of accuracy.

A last note is about the π_0 in Storey's q-value method which is used to take into account the multiple testing problem. The critical part of q-value method of controlling the false discovery rate is to give a good estimate of π_0, the proportion of the true null hypotheses among all the tests. The current method we used in this study is based on the algorithm suggested by Storey et al. (2003) which assumes the distribution of p-value from each test is uniform over the (0,1) interval. However, Yang (2004) pointed out that if the p-values were not uniformly distributed, the

power of the q-value method may decrease [73]. He suggested to compute a weighted average of π_0 from the distribution of the raw p-values which are greater than a threshold (say 0.4). Thus, it gives a better control and more robust estimate of π_0. It is also worthwhile to try some other good multiple-testing methods as well.

Chapter A

Appendix

A.1 Derivation of $\boldsymbol{S, V}$

The entries of the matrices $\boldsymbol{S, V}$ are derived by repeated differentiation of the equation (3.7) based on the fact that $\int dG(t) = 1$ and $\int \omega_j(t) dG(t) = 1$, $j = 1, \ldots, q$. This is an extension of Fokianos et al. (2001) [13] to the vector case of the tilt function $\boldsymbol{h}(x)$.

First define

$$\nabla \equiv \left(\frac{\partial}{\partial \alpha_1}, ..., \frac{\partial}{\partial \alpha_q}, \frac{\partial}{\partial \boldsymbol{\beta}_1} ..., \frac{\partial}{\partial \boldsymbol{\beta}_q} \right)' \tag{A.1}$$

Then $E[\nabla l(\alpha_1, ..., \alpha_q, \boldsymbol{\beta}_1, .., \boldsymbol{\beta}_q)] = \mathbf{0}$. To obtain the score second moments it is convenient to define $\rho_m \equiv 1,\ w_m(t) \equiv 1$,

$$E_j[\boldsymbol{h}(t)] \equiv \int \boldsymbol{h}(t) w_j(t) dG(t) \tag{A.2}$$

and,

$$A_0(j, j') \equiv \int \frac{w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \tag{A.3}$$

$$\boldsymbol{A}_1(j, j') \equiv \int \frac{\boldsymbol{h}(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \tag{A.4}$$

$$\boldsymbol{A}_2(j, j') \equiv \int \frac{\boldsymbol{h}(t) \boldsymbol{h}'(t) w_j(t) w_{j'}(t) dG(t)}{1 + \sum_{k=1}^{q} \rho_k w_k(t)} \tag{A.5}$$

for $j, j' = 1, ..., q$. Then, the entries in

$$\boldsymbol{V} \equiv Var \left[\frac{1}{\sqrt{n}} \nabla l(\alpha_1, ..., \alpha_q, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q) \right] \tag{A.6}$$

are,

$$\frac{1}{n}Var\left(\frac{\partial l}{\partial \alpha_j}\right) = \frac{\rho_j^2}{1+\sum_{k=1}^q \rho_k}\{A_0(j,j) - \sum_{r=1}^m \rho_r A_0^2(j,r)\} \quad (A.7)$$

$$\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \alpha_{j'}}\right) = \frac{\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\{A_0(j,j') - \sum_{r=1}^m \rho_r A_0(j,r)A_0(j',r)\} \quad (A.8)$$

$$\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_j}\right) = \frac{\rho_j^2}{1+\sum_{k=1}^q \rho_k}\{A_0(j,j)E_j[\boldsymbol{h}'(t)] - \sum_{r=1}^m \rho_r A_0(j,r)\boldsymbol{A}_1'(j,r)\} \quad (A.9)$$

$$\frac{1}{n}Cov\left(\frac{\partial l}{\partial \alpha_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_{j'}}\right) = \frac{\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\{A_0(j,j')E_{j'}[\boldsymbol{h}'(t)] - \sum_{r=1}^m \rho_r A_0(j,r)\boldsymbol{A}_1'(j',r)\} \quad (A.10)$$

$$\begin{aligned}\frac{1}{n}Cov\left(\frac{\partial l}{\partial \boldsymbol{\beta}_j}, \frac{\partial l}{\partial \boldsymbol{\beta}_{j'}}\right) = \frac{\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\{&-\boldsymbol{A}_2(j,j') + E_j[\boldsymbol{h}(t)]\boldsymbol{A}_1'(j,j') \\ &+ \boldsymbol{A}_1(j,j')E_{j'}[\boldsymbol{h}'(t)] \\ &- \sum_{r=1}^m \rho_r \boldsymbol{A}_1(j,r)\boldsymbol{A}_1'(j',r)\} \\ &+ \frac{1}{n}\sum_{i=1}^{n_j}\sum_{k=1}^{n_{j'}} Cov[\boldsymbol{h}(\epsilon_{ji}), \boldsymbol{h}(\epsilon_{j'k})]\end{aligned} \quad (A.11)$$

The last term is 0 for $j \neq j'$ and $(n_j/n)Var[\boldsymbol{h}(\epsilon_{j1})]$ for $j = j'$.

Next, as $n \to \infty$,

$$-\frac{1}{n}\nabla\nabla' l(\alpha_1, ..., \alpha_q, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q) \to \mathbf{S} \quad (A.12)$$

where $\mathbf{S}$ is a $q(1+p) \times q(1+p)$ matrix with entries corresponding to $j, j' = 1, ..., q$,

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j^2} \to \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]w_j(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad (A.13)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial \alpha_j \alpha_{j'}} \to \frac{-\rho_j \rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{w_j(t)w_{j'}(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad (A.14)$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial\alpha_j\partial\boldsymbol{\beta}_j'} \rightarrow \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]w_j(t)\boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad \text{(A.15)}$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial\alpha_j\partial\boldsymbol{\beta}_{j'}'} \rightarrow \frac{-\rho_j\rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{w_j(t)w_{j'}(t)\boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad \text{(A.16)}$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j'} \rightarrow \frac{\rho_j}{1+\sum_{k=1}^q \rho_k}\int \frac{[1+\sum_{k\neq j}^q \rho_k w_k(t)]w_j(t)\boldsymbol{h}(t)\boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad \text{(A.17)}$$

$$-\frac{1}{n}\frac{\partial^2 l}{\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_{j'}'} \rightarrow \frac{-\rho_j\rho_{j'}}{1+\sum_{k=1}^q \rho_k}\int \frac{w_j(t)w_{j'}(t)\boldsymbol{h}(t)\boldsymbol{h}'(t)}{1+\sum_{k=1}^q \rho_k w_k(t)}dG(t) \quad \text{(A.18)}$$

It should be noted that, due to profiling, the matrix $\boldsymbol{S}$ is not the usual information matrix although it plays a similar rote.

Thus when the density ratio model (3.1) holds for the true parameters α_0 and $\boldsymbol{\beta}_0$, it follows under the regularity condition that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are both consistent and asymptotically normal as in (3.10) (see Sen and Singer 1993, chapter 5) [56],

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0 \\ \\ \hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0 \end{pmatrix} \Rightarrow \mathrm{N}(\mathbf{0},\boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}=\boldsymbol{S}^{-1}\boldsymbol{V}\boldsymbol{S}^{-1}$

A.2 Simplified Semiparametric Test Statistics for Binary Data

If the data are 0-1 binary data, such as cancer or no cancer, we can simplify the likelihood and obtain closed forms for the parameter estimates. Recall:

N_G: The combined sample size for the whole study region.

n_G: The number of cases in the whole study region.

N_Z: The sample size within the scan window.

n_G: The number of cases within the scan window.

t_i: The ith observation from the combined sample in the whole study region, $i = 1, \ldots, N_G$.

ρ_1: The relative sample size, which is equal to $N_Z/(N_G - N_Z)$.

x_{Zj}: The jth observation within the scan window Z, $j = 1, \ldots, N_Z$.

For binary data, choose the tilt function $h(x) = x$. Then the profile log-likelihood with parameters α_1 and β_1 is

$$\begin{aligned} \ell(\alpha_1, \boldsymbol{\beta}_1) &= -\sum_{i=1}^{N_G} \log[1 + \rho_1 e^{\alpha_1 + \beta_1 t_i}] + \sum_{j=1}^{N_Z} (\alpha_1 + \beta_1 x_{Zj}) \\ &= -(N_G - n_G) \cdot \log\left[1 + \frac{N_Z}{N_G - N_Z}\, e^{\alpha_1}\right] \\ &\quad - n_G \cdot \log\left[1 + \frac{N_Z}{N_G - N_Z}\, e^{\alpha_1 + \beta_1}\right] + N_Z \cdot \alpha_1 + n_Z \cdot \beta_1 \end{aligned} \tag{A.19}$$

The resulting maximum likelihood estimators are,

$$\begin{cases} \hat{\alpha}_1 &= \log\left(\frac{N_Z - n_Z}{N_Z}\right) - \log\left(\frac{(N_G - N_Z) - (n_G - n_Z)}{N_G - N_Z}\right) \\ \hat{\beta}_1 &= \log\left(\frac{n_Z}{N_Z - n_Z}\right) - \log\left(\frac{n_G - n_Z}{(N_G - N_Z) - (n_G - n_Z)}\right) \\ e^{\hat{\alpha}_1 + \hat{\beta}_1} &= \frac{n_Z/N_Z}{(n_G - n_Z)/(N_G - N_Z)} \end{cases} \tag{A.20}$$

Apparently, $\beta_1 = 0$ implies $\alpha_1 = 0$ and $e^{\alpha_1+\beta_1} = 1$, which means the relative rates inside and outside the scan window are equal.

By equation (3.15), the χ_1 test statistic can be simplified to

$$\chi_1 = N_G \cdot \hat{\beta}_1 \cdot \left\{ \frac{\rho_1}{(1+\rho_1)^2} \cdot \left[\frac{n_G - n_Z}{N_G - N_Z} \cdot (1 - \frac{n_G - n_Z}{N_G - N_Z}) \right] \right\} \cdot \hat{\beta}_1 \qquad \text{(A.21)}$$

where $\rho_1 = \frac{N_Z}{N_G - N_Z}$ and $\hat{\beta}_1$ is as in equation (A.20).

By equation (3.16), the likelihood ratio test statistic can be simplified to

$$\begin{aligned} LR &= 2\log\left[(\frac{n_Z}{N_Z})^{n_Z} \cdot (\frac{n_G - n_Z}{N_G - N_Z})^{n_G - n_Z} \cdot (\frac{N_Z - n_Z}{N_Z})^{N_Z - n_Z} \right] \\ &\quad + 2\log\left[(\frac{(N_G - N_Z) - (n_G - n_Z)}{N_G - N_Z})^{(N_G - N_Z) - (n_G - n_Z)} \right] \\ &\quad - 2\log\left[(\frac{n_G}{N_G})^{n_G} \cdot (\frac{N_G - n_G}{N_G})^{N_G - n_G} \right] \\ &= 2\log[\text{Kulldoff's Bernoulli scan stat. as in equation (2.1)}] \qquad \text{(A.22)} \end{aligned}$$

If the tilt function $\boldsymbol{h}(x)$ is chosen as (x, x^2), the parameter β_{12}, which is associated with the x^2 term, is actually 0. Notice that the normal equation of the likelihood for β_{11}, which is corresponding to the x term, is identical with the normal equation for β_{12}. Thus the x^2 term is confounded with the x, and β_{12} is not estimable. Therefore, only α_1 and β_{11} are estimated, and the final results are still the same as in the situation where $h(x) = x$.

Bibliography

[1] Alexander, F., Cartwright, R., McKinney, P.A., and Ricketts T.J. (1990). Investigation of spatial clustering of rare diseases: childhood malignancies in North Humberside. *Journal of Epidemiology and Community Health*, 44, 39-46.

[2] Barlow R.E., Bartholomew D. J., Bremner J. M., Brunk H. D. (1974). Statistical Inference under Order Restrictions. Wiley, New York.

[3] Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B, Vol. 57, No. 1, 289-300.

[4] Bonetti, M. and Pagano, M. (2005). The interpoint distance distribution as a descriptor of point patterns: An application to cluster detection. *Statistics in Medicine* (in press).

[5] Colt, J.S. and Blair, A. (1998). Parental Occupational Exposures and Risk of Childhood Cancer, *Environmental Health Perspectives Supplements*, 106, No. S3, 909-925.

[6] Cuzick J. and Edwards R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society*, series B, 52: 73-104.

[7] Anna Ditkoff, "Murder Ink"., Baltimore City Paper (January 11, 2006)

[8] Morgan Quitno's America's Safest Cities, "City Crime Rankings by Population Group". `http://www.morganquitno.com/cit06pop.htm#500,000+`

[9] CNN/Morgan Quitno's America's Safest Cities, "Top 25: Most dangerous and safest cites". `http://money.cnn.com/2006/10/30/real_estate/Most_dangerous_cities/index.htm`

[10] Dwass M. (1957). Modified Randomization tests for Nonparamertic hypotheses. *The Annuals of Mathematical Statistics*, 28, 181-187.

[11] Dykstra R., Kochar S., Robertson T. (1995). Inference for Likelihood Ratio Ordering in the Two-Sample Problem. *Journal of the American Statistical Association*, Vol. 90, No. 431, 1034-1040.

[12] Eck J.E., Chainey S., Cameron J.G., Leitner M., Wilson R.E. (2005). Mapping Crime: Understanding Hot Spots. *NIJ special report, U.S. Department of Justice, Office of Justice Programs, National Institute of Justice*, Aug, 2005.

[13] Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43, 56-65.

[14] Fokianos K. (2006). Density Ratio Model Selection. *Journal of Statistical Computations and Simulation*. To appear.

[15] Fokianos K. and Kaimi I., (2006). On the effect of misspecifying the density of ratio model. *Annals of the Institute of Statistical Mathematics*, Vol. 58, No. 3, 475-497.

[16] Gagnon R. (2005), *Certain Computational Aspects of Power Efficiency and of State Space Models*. Ph.D Dissertation, Department of Mathematics, University of Maryland, College Park, Maryland.

[17] Glaz J. and Naus J. (1991). Tight bounds for scan statistics probabilities for discrete data. *The Annals of Applied Probability*, Vol. 1, No., 306-318.

[18] Glaz J., Naus J., and Wallenstein S. (2001). *Scan statistics*. Springer, New York.

[19] Glaz J. and Zhang Z. (2004). Multiple Window Discrete Scan Statistics. *Journal of Applied Statistics*, Vol. 31, Issue 8, 967-980.

[20] Glaz J. and Zhang Z. (2006). Maximum scan score-type statistics *Statistics and Probability Letters*, Volume 76, Issue 13, 1316-1322.

[21] Hjalmars U., Kulldorff M., Gustafsson G., and Nagarwalla N. (1996). Childhood leukemia in Sweeden using GIS and a spatial scan statistics for cluster detection. *Statistics in Medicine*, 15, 707-715.

[22] Huang L., Kulldorff M., Gregorio D. (2007). A Spatial Scan Statistic for Survival Data. *Biometrics* 63 (1), 109-118.

[23] Jung I., Kulldorff M., Klassen A.C. (2007). A spatial scan statistic for ordinal data. *Statistics in Medicine*, Volume 26, Issue 7, 1594-1607.

[24] Kay, R., and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data. *Biometrika*, 74, 495-501.

[25] Kedem, B., Wolff, D.B., and Fokianos, K. (2004). Statistical Comparison of Algorithms. *IEEE Transactions on Instrumentation and Measurement*, 53, 770-776.

[26] Kedem B., Wen S. (2007). Semi-parametric Cluster Detection. *Journal of Statistical Theory and Practice*, inaugural issue, 1, 49-72.

[27] Keziou, A. and Leoni-Aubin, S. (2005). Test of homogeneity in semiparametric two-sample density ratio models. *Comptes Rendus de l'Academie des Sciences, Paris*, Ser. I 340, 905-910.

[28] Klassen A.C., Kulldorff M., Curriero F.C. (2005). Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics 2005*, 4:1.

[29] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics–Theory and Methods*, 26, 1481-1496.

[30] Kulldorff M. (1999). Spatial Scan Statistics: Models, Calculations, and Applications. *Scan Statistics and Applications*, edited by Joseph Glaz and N. Balakrishnan, Birkhäuser, Boston.

[31] Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic, *Journal of the Royal Statistical Society* A, 164, part 1, 61-72.

[32] Kulldorff, M., Tango T. Park P.J. (2003), Power comparison for disease clustering tests, *Computational Statistics & Data Analysis*, 42, 665-684.

[33] Kulldorff M., Heffernan R., Hartman J., Assunção R.M., Mostashari F. (2005). A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine*, 2: 216-224.

[34] Kulldorff M., Huang L., Pickle L., Duczmal L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, Volume 25, Issue 22, 3929-3943.

[35] Normal Model: Kulldorff M, et al., (2006), manuscript in preparation.

[36] Kulldorff M., Mostashari F., Duczmal L., Yih W.K., Kleinman K., Platt R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, Volume 26, Issue 8, 1824-1833

[37] LaFree, G., Morris, N., Dugan, L. and Fahey, S. (2006). Identifying Global Terrorist Hot Spots. *The Psychology of Terrorism*, IOS Press.

[38] Lawson A.B. (Editor), Biggeri A. (Editor), Böhning D. (Editor), Lesaffre E. (Editor), Jean-Fran Viel J. (Editor), Bertollini R. (Editor). (1999). *Disease Mapping and Risk Assessment for Public Health*, WILEY.

[39] LeBeau J.L. (1999). Demonstrating The Analytical Utility of GIS for Police Operations A Final report to Grant: NIJ 97-LB-VX-K010.

[40] Lu, G. (2007). Asymptotic Theory for Multiple-Sample Semiparametric Density Ratio Models and Its Application To Mortality Forecasting. *Ph.D. dissertation*, University of Maryland.

[41] McKinney P.A., Alexander F.E., Cartwright R.A., Parker L. (1991). Parental occupations of children with leukemia in west Cumbria, North Humberside, and Gateshead. *British Medical Journal*, 302, 681-687.

[42] Modarres R. and Patil G.P. (2006). Hotspot Detection with Bivariate Data. *Journal of Statistical Planning and Inference* , (S.N. Roy Centennial Volume) (in press).

[43] 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.

[44] Naus J.I. and Wartenberg D. (1997). A double-scan statistic for clusters of two types of events. *Journal of the American Statistical Association*, 92: 1105-1113.

[45] Naus J.I. and Wallenstein S. (2004). Multiple Window and Cluster Size Scan Procedures. *Journal Methodology and Computing in Applied Probability*, Vol. 6, No 4, 389-400.

[46] Newell G.G. (1963). Distribution for the samllest distance between any pair of Kth nearest-neighbor random points on a line. In Rosenblatt, M. ed., *Time series Analysis, Processdings of a Conference Held at Brown University.* Wiley, New York, pp. 89-103.

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

[48] Patil G.P., Rathbun S.L., Acharya R., Patankar P., Modarres R. (2005). Upper Level Set Scan Statistic System for Detecting Arbitrarily Shaped Hotspots for Digital Governance. *Proceedings of the 2005 national conference on Digital government research*; Vol. 89, 281-282.

[49] Pickle, L.W., Feuer, E.J., and Edwards, B.K. (2003). *U.S. Predicted Cancer Incidence, 1999: Complete Maps by County and State From Spatial Projection*

Models, National Cancer Institute, Cancer Surveillance Monograph No. 5, NIH Publication No. 03-5435.

[50] Pozdnyakov V., Glaz J., Kulldorff M., and Steele M. (2004). A martingale approach to scan statistics. *Annals of the Institute of Statistical Mathematics*, Volume 57, Number 1, 21-37.

[51] Qin, J. (1993). Empirical likelihood in biased sampling problems. *The Annals of Statistics*, 21, 1182-1186.

[52] Qin, J., and Lawless, J.F. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

[53] Qin, J., and Zhang, B. (1997). A goodness of fit test for logistic regression models based on case-control data. *Biometrika*, 84, 609-618.

[54] Reiner A., Yekutieli D., Yoav Benjamini Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, Vol. 19, No. 3, 368-375.

[55] Thomas Rich (1999). Mapping the Path to Problem Solving. *National Institute of Justice Journal*, October 1999.

[56] Sen P.K. and Singer J.M. (1993). *Large sample method in statistics, An introduction with Applications*. Chapman and Hall/CRC, London.

[57] Shmueli G., Thomas M.P., Kadane J.B., Borle S., Boatwright P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society*, Series C (Applied Statistics), 54 (1), 127-142.

[58] Shmueli, G., and Fienberg, S. E. (2006). Current and Potential Statistical Methods for Monitoring Multiple Data Streams for Bio-Surveillance. *Statistical Methods in Counter-Terrorism: Game Theory, Modeling, Syndromic Surveillance, and Biometric Authentication*, Eds: A Wilson, G Wilson, and D H Olwell, Springer, New York.

[59] Storey J.D. and Tibshirani R. (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report 2001-28, Department of Statistics, Stanford University.* (NOTE: This paper was reconceived and rewritten into Storey and Tibshirani (2003) PNAS, which is listed below.)

[60] Storey J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, Series B, 64: 479-498.

[61] Storey J.D., Tibshirani R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences*, 100, 9440-9445.

[62] Storey J.D., (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31, 2013-2035.

[63] Storey J.D., Taylor J.E., Siegmund D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society*, Series B, 66: 187-205.

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

[65] Tango T. and Takahashi K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11.

[66] Tsai C., Hsueh M., Chen J., (2003). Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data. *Biometrics*, Vol. 59, No. 4, 1071-1081.

[67] Wallenstein S. and Neff N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medice*, 6(2): 197-207.

[68] Waller, L.A. and Lawson, A.B. (1995). The power of focused tests to detect disease clustering. *Statistics in Medicine*, 14, 2291-2308.

[69] Weisburd D. and Braga A.A. (2006). Police Innovation: Contrasting Perspectives (Cambridge Studies in Criminology). Cambridge University Press, UK.

[70] Wen S. and Kedem B. (2007). Power Study of a Semi-parametric Cluster Detection Method. Submitted to *Journal of Computational Statistics and Data Analysis*.

[71] Wikipedia on "Montgomery County, Maryland", `http://en.wikipedia.org/wiki/Montgomery_County/`

[72] Wikipedia on "Fairfax County, Virginia", `http://en.wikipedia.org/wiki/Fairfax_County/`

[73] Yang X. (2004). Qvalue Methods May not Always Control False Discovery Rate in Genomic Applications, *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, 556- 557.