Distribution Agreement

Signature:

Date:

Statistical Performance of Spatial Systems

By

Yuemei Wang

Doctor of Philosophy

Biostatistics

Lance A. Waller, Ph.D.
Advisor

DuBois Bowman, Ph.D.
Committee Member

James W. Buehler, MD.
Committee Member

Qi Long, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Statistical Performance of Spatial Systems

By

Yuemei Wang

B.S., NanKai University, 1995

M.S., Emory University, 2007

Advisor: Lance A. Waller, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2009

Abstract

Statistical Performance of Spatial Systems

By Yuemei Wang

Detection of disease outbreaks is a crucial issue in public health. Therefore, we want statistical methods to evaluate the accuracy and reliability of proposed detection systems.

Furthermore, detecting outbreaks in space is very challenging as the shapes and locations of outbreak clusters of disease can be unpredictable. In this thesis, we use area under the receiver operating characteristic (ROC) curve to evaluate statistical performance of several proposed spatial detection systems. Specifically, we assess spatial statistical performance of two spatial scan statistics, and their applications to cardiac birth defect data from Santa Clara County, California. The results reveal SaTScan performs better if the cluster is compact, and Upper Level Set approaches offer improved performance when clustering is irregularly shaped.

We also investigate performance of cluster detection methods when adjusting for covariates via Generalized Additive Models (GAM). We apply GAMs to archaeological data from Black Mesa, Arizona to identify clusters of early versus late Anasazi settlement sites when adjusting for exposure to rivers around those sites. Furthermore, we evaluate spatial variations in power to detect different levels of clustering when clusters are allowed to occur at different locations within this application. We compare the GAM results and performance of the GAM methodology with those based on kernel density estimation of the early-to-late relative risk surface.

Finally, we assess spatial performance of detection systems using decision fusion theory for the situation where a detection system can be comprised of a few expensive, precise detectors and many inexpensive, imprecise detectors. The performance of a system depends not only on the total allowable cost for the system, but also the performance of each individual detector, as well as the balance between expensive, precise and inexpensive, imprecise detectors. We quantify how, if we improve the performance of imprecise detectors even slightly, the performance of the resulting system improves dramatically. In addition, we show that lower-cost systems can perform as well as or better than systems expending the full allowable cost. These results indicate the need for careful calculation and computation to identify an optimal system, especially for systems comprised of small numbers of components.

Statistical Performance of Spatial Systems

By

Yuemei Wang
B.S., NanKai University, 1995
M.S., Emory University, 2007

Advisor: Lance A. Waller, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2009

Acknowledgements

I am extremely grateful to have had the opportunity to be a part of Emory University and Department Of Biostatistics and Bioinformatics. I would like to extend my deep gratitude to the many people who were instrumental in helping me to complete this body of work.

First, I would like to express my deepest gratitude to my advisor and committee chair, Dr. Waller. His thoughtful guidance, key insights, endless patience, critical assessments, and enduring friendship provided me with a positive environment in which to do research. I am fortunate to have so many fond memories and honored to have worked with him.

I would also like to thank my committee members, Dr. Bowman, Dr. Buehler and Dr. Qi, who have generously given their time and expertise to better my work. Thanks to my friends and the department faculty and staff for making my time at Emory University memorable. I also would like to extend my gratitude to the Department of Biostatistics and Bioinformatics for the scholarships and assistantships awarded to me. I am profoundly grateful to Dr. Patil from Penn State University for providing me with the opportunity to study and learn as part of his research group, and continuous support after I transferred to Emory University.

Finally, I could not have done this without the love and support of my family. I am grateful to my mother for instilling in me values and ideals I hold dear. I would especially like to thank my husband and friend, Shixin, for understanding and indulging me during this time. Special thanks go to my boys, Jesse and Nathan, for accepting the limits placed on our time together during the writing of this thesis. They are my inspiration and deserves much of the credit for motivating me through my research. To them, I dedicate this thesis.

My love and best of wishes to all of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Introduction

Detection of disease outbreaks is a crucial issue in public health. If a communicable disease outbreak is detected within a population at a very early stage, mortality and morbidity may be significant decreased by prompt response of health officials. Failure or delay in detecting an outbreak could jeopardize more people's lives. However, false alarms could cause public panic and waste resources. Therefore, accurately detecting an outbreak in a timely manner is very important.

Statistical methods are useful in outbreak detection. The software program SaTscan has been used for monitoring West Nile virus (18) and New York City public health surveillance (63). Autoregressive analysis has been applied to disease outbreak detection system in the Washington DC area (50). Survey analysis and case control studies have contributed to detecting Salmonella infections in Oregon in 1996 using CDC's FoodNet surveillance system(17). Evaluation of the accuracy and reliability of statistical detection algorithms and systems is critical to accurate disease surveillance. Bravata et. al. reviewed 115 systems involving surveillance systems, however, only three systems have evaluated sensitivity and specificity as a measure of performance of the systems (14).

In many instances, detection of disease outbreaks involves spatial analysis. Our goal is to derive methods to evaluate the statistical performance of spatial surveillance systems. We are particularly interested in methods that assess performance spatially, i.e., we want maps of how well methods detect outbreaks at different locations.

1.2 Public Health Surveillance

Surveillance is the action watching over activities. A surveillance system includes data collection, data analysis and dissemination of results of the analysis (73). Specific to our application, the US Centers for Diseases Control and Prevention (CDC) defines public health surveillance as "the ongoing, systematic collection, analysis, and interpretation of data (e.g., regarding agent/hazard, risk factor, exposure, health event) essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those responsible for prevention and control" (73). Public health surveillance can be used to detect new outbreaks of a disease, to provide information about scope, size or history of a disease or an epidemic, and to evaluate the impact of health interventions (61).

The history of public health surveillance goes back at least to the fourteenth century and the Black Death in Europe, where the disease was kept under control by 40-day detention of travellers from infected areas (68). After September 11, 2001, the objectives of public health surveillance focused not only on naturally occurring diseases, but also shifted to include potential bioterrorist attacks (1; 23; 40; 59; 70). Therefore, early detection has become an increasingly important criterion within public health surveillance. With the rapid development of geographic information systems and availability of locations of cases, spatial analysis also is becoming a routine part of surveillance including the detection and evaluation of incident clusters.

Public health surveillance can be passive or active. A passive surveillance system relies on hospitals, health care providers, laboratories, pharmacies, or other reporters to provide case information to the system. In contrast, an active surveillance system will contact clinics, hospitals, or other resources for required information rather than wait for reports. Usually, an active approach will have higher quality of data than a passive system, but the cost is often higher (3).

There are two types of evaluation commonly related to public health surveillance systems. The first is to decide whether a particular health event should be under surveillance, and the second is to evaluate an existing surveillance system to check its usefulness, effectiveness, simplicity, flexibility, acceptability, cost and efficiency. Here we concentrate on the second type and focus on quantitative attributes of a public health surveillance system, such as sensitivity, specificity, and predictive value positive (PVP), each of which focuses on the extent to which true cases are among those identified by the system (61).

1.3 Clustering

Spatial surveillance often involves a search for disease clusters (2; 12). A disease cluster refers to areas of abnormal observations, such as those with a higher incidence rate, compared to other areas in a particular time period, in space or both (12; 61). Methods for detecting disease clusters provide tools for exploring data, but causal effects are typically drawn by further research (28).

The public is very concerned with disease clustering, and over 1,000 cancer cluster reports are filed each year in USA, and more than one cluster is reported per week in the UK (10; 12; 58). One goal of disease cluster analysis is to test whether a suspect area is a cluster with statistical elevated risk. For example, in 1981 a leak of methyl chloroform was detected in Santa Clara County, California. This report was accompanied by reports of increased numbers of congenital cardiac anomalies and adverse pregnancy outcomes from residents in those areas where they might be exposed to the contaminated water (43; 71). It is important to note that further evidence beyond a simple spatial pattern analysis is required to support exposure to the solvent as the definitive cause of this cluster.

Another goal of disease cluster analysis is to identify clusters independently of putative causes (25), then conduct further investigation to find the cause of these “hot spots”, but such goals are subject to debate (10; 12). Rothman has warned that little scientific value has been gained by retrospective cluster analysis (41; 65). Rather than respond to each report individually, health departments often build on-going active cluster detection within proactive surveillance systems on a state or national level. With the rapid development of geographic information systems, software tools, and statistical methods, one can identify patterns, and further pursue in-depth research. For example, an ongoing surveillance system could provide notification when the median relative interval of certain disease or the mean time interval of, say, five cases reach a cut off value in one county (62).

1.4 Measure of Performance

To place these concepts in a statistical framework, suppose the null hypothesis is that the disease rate is the same across the region, and the alternative hypothesis is that some areas have higher incidence rate than the other areas. That is,

$$H_0 : R_i = \lambda \ \forall i \in I$$
$$H_1 : R_i > \lambda \text{ for some } j \in I$$

where R_i is the disease rate at location i, and λ is the average disease rate across the region.

In this thesis, we investigate the statistical performance of tests based on such hypotheses. The following are some ways to measure performance of an individual test.

1.4.1 Sensitivity

Sensitivity is the probability of detecting an outbreak given the outbreak truly exits. Let T denote the binary outcome of a detection test (present/absent), then Sensitivity $= \text{Prob}(T = \text{present}|H_1)$. For normally distributed tests, we assume that under the null hypothesis T follows a standard normal distribution, that is, $T \sim N(\mu_0, \sigma_0^2) = N(0,1)$, and under the alternative hypothesis T follows a normal distribution $T \sim N(\mu_i, \sigma_i^2)$ at location i, where $\mu_i \neq 0$ and/or $\sigma_i \neq 1$. In addition, if t^* is the cut point of the test, we declare the outbreak is present when we observe $T > t^*$. Also let $Z \sim N(0,1)$ and Φ denote the cumulative distribution function of Z, then the following basic result holds:

$$\begin{aligned}
\text{Sensitivity} &= \text{Prob}(T = \text{present}|H_1) \\
&= \text{Prob}(T > t^*|\mu = \mu_i, \sigma^2 = \sigma_i^2) \\
&= \text{Prob}\left(\frac{T-\mu_i}{\sigma_i} > \frac{t^*-\mu_i}{\sigma_i}\right) \\
&= \text{Prob}\left(Z > \frac{t^*-\mu_i}{\sigma_i}\right) \\
&= 1 - \text{Prob}\left(Z < \frac{t^*-\mu_i}{\sigma_i}\right) \\
&= 1 - \Phi\left(\frac{t^*-\mu_i}{\sigma_i}\right) \qquad (1.1)
\end{aligned}$$

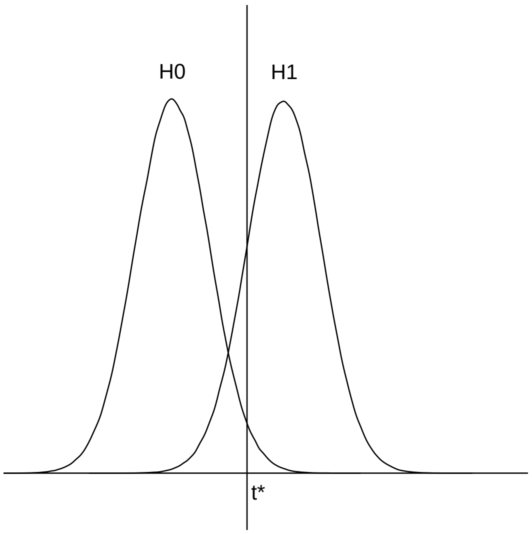


Figure 1.1: Sensitivity and Specificity

1.4.2 Specificity

Specificity is the probability of declaring no outbreak given the outbreak is truly absent. Similar to our discussion of sensitivity, for a binary outcome of a normally distributed tests, the following result holds:

$$\begin{aligned} \text{Specificity} &= \text{Prob}(T = \text{absent}|H_0) \\ &= \text{Prob}(T < t^*|\mu = 0, \sigma^2 = 1) \\ &= \text{Prob}(Z < t^*) \\ &= \Phi(t^*) \end{aligned} \tag{1.2}$$

As shown in Figure 1.1, sensitivity is the area on the right side of t^* under H_1, and specificity is the area on the left side of t^* under H_0.

1.4.3 PPV and NPV

Positive predictive value (PPV) is the probability that an outbreak truly exists given a positive test result. Negative predictive value (NPV) is the probability that the outbreak is absent given a negative test result. However, PPV and NPV not only depend on the background performance of a test, but also on the prevalance of the disease, i.e., $P(H_1)$. The relationships between PPV, NPV and sensitivity, specificity are shown in equations (1.3) and (1.4).

$$\begin{aligned}
\text{PPV} &= \text{Prob}(H_1|T = \text{present}) \qquad (1.3) \\
&= \frac{\text{Prob}(H_1, T = \text{present})}{\text{Prob}(T = \text{present})} \\
&= \frac{\text{Prob}(T = \text{present}|H_1) \times \text{Prob}(H_1)}{\text{Prob}(T = \text{present})} \\
&= \frac{\text{Prob}(T = \text{present}|H_1) \times \text{Prob}(H_1)}{\text{Prob}(T = \text{present}|H_1) \times \text{Prob}(H_1) + \text{Prob}(T = \text{present}|H_0) \times \text{Prob}(H_0)} \\
&= \frac{\text{Sensitivity} \times \text{Prob}(H_1)}{\text{Sensitivity} \times \text{Prob}(H_1) + \text{Specificity} \times (1 - \text{Prob}(H_1))}
\end{aligned}$$

$$\begin{aligned}
\text{NPV} &= \text{Prob}(H_0|T = \text{absent}) \qquad (1.4)\\
&= \frac{\text{Prob}(H_0, T = \text{absent})}{\text{Prob}(T = \text{absent})}\\
&= \frac{\text{Prob}(T = \text{absent}|H_0) \times \text{Prob}(H_0)}{\text{Prob}(T = \text{absent})}\\
&= \frac{\text{Prob}(T = \text{absent}|H_0) \times \text{Prob}(H_0)}{\text{Prob}(T = \text{absent}|H_0) \times \text{Prob}(H_0) + \text{Prob}(T = \text{absent}|H_1) \times \text{Prob}(H_1)}\\
&= \frac{(1 - \text{Specificity}) \times \text{Prob}(H_0)}{(1 - \text{Specificity}) \times (1 - \text{Prob}(H_1)) + (1 - \text{Sensitivity}) \times \text{Prob}(H_1)}
\end{aligned}$$

1.4.4 ROC and AUC

Receiver operating characteristic (ROC) curves are popular summaries of statistical performance and combine information related to both sensitivity and specificity. In particular, the ROC curve plots sensitivity vs. false alarm rate (1- specificity), across a range of potential critical values.

Figure 1.2 shows simulated ROC curves with $T \sim N(0, 1)$ under H_0 and $T \sim N(1, 1)$, $T \sim N(2, 1)$ or $T \sim N(4, 1)$under H_1, where the cut points range from 5 to -5 with increment of -0.01. When the cut point is set very high, the false alarm rate is low, but the probability of catching a true positive is also low as shown on the left of ROC curve. On the other hand, if the cut point is set very low, we will identify most outbreaks, but the false alarm rate will be high as shown on the right of the ROC curve. Therefore, a reasonable cut point should be used with acceptable false alarm rate and sensitivity level. A perfect test should have false alarm rate equal to zero, and sensitivity equal to one. On the other hand, a noninformative test has a

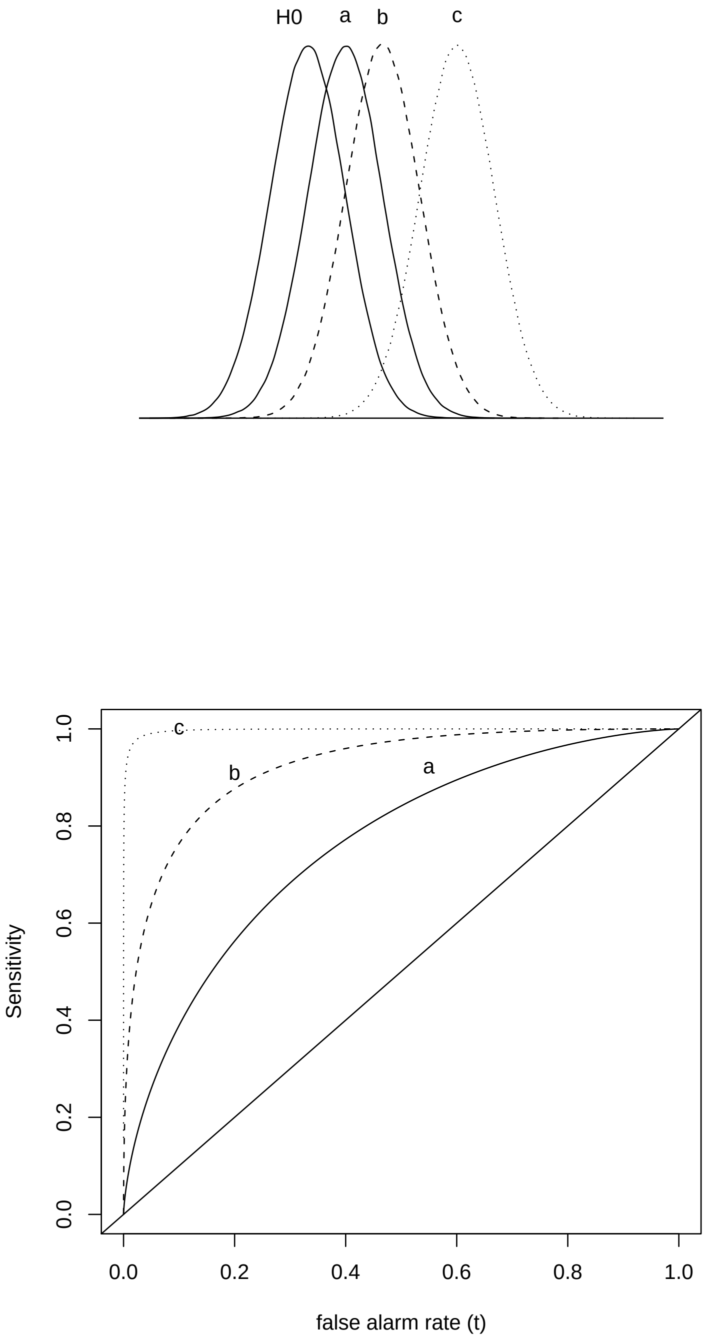


Figure 1.2: ROC curves

false alarm rate equal to sensitivity, i.e., an ROC curve lying on the 45 degree line. Therefore, the closer the ROC curve is to the diagonal, the less useful the test is; and the more steeply the curve moves up to the upper left corner of the ROC plot, the more useful the test is. As a result, the area under the ROC curve (AUC) is a good indicator of the performance of a test. The closer the AUC is to 1.0, the better the test, and the closer the AUC is to 0.5, the worse the test. If the AUC is 1.0, the test is perfect.

Assume $T \sim N(0, 1)$ under H_0 and $T \sim N(\mu_i, \sigma_i^2)$ under H_1, Figure 1.3 shows that AUC decreases as σ_i increases when μ_i is held constant (right panel), and AUC increases as μ_i increases and σ_i is held constant (left panel). Figure 1.4 shows the contour plot of AUC with normally distributed test statistics as both σ_i and μ_i change. In summary, the more overlap between the distribution of T under the null hypothesis and the alternative hypothesis, the smaller the AUC is.

The ROC curve and AUC have seen broad application in many areas, including, but not limited to, medical decision making (53) and signal processing (72).

1.5 Summary of Thesis and Primary Contributions

In the following chapters, we first review cluster detection methods, then describe two data sets to be used later in our research. Detecting outbreaks in space is very challenging as the shapes and locations of outbreak clusters of disease can be unpredictable. In chapter four, we assess spatial statistical performance of two spatial scan statistics (the SaTScan and the Upper Level Set stastistics), and their applications to cardiac birth defect data from Santa Clara County, California. In chapter five, we investigate performance of cluster detection methods when adjusting for covariates via Generalized Additive Models (GAM). We apply GAMs to archaeological data from Black Mesa, Arizona to identify clusters of early versus late Anasazi settlement

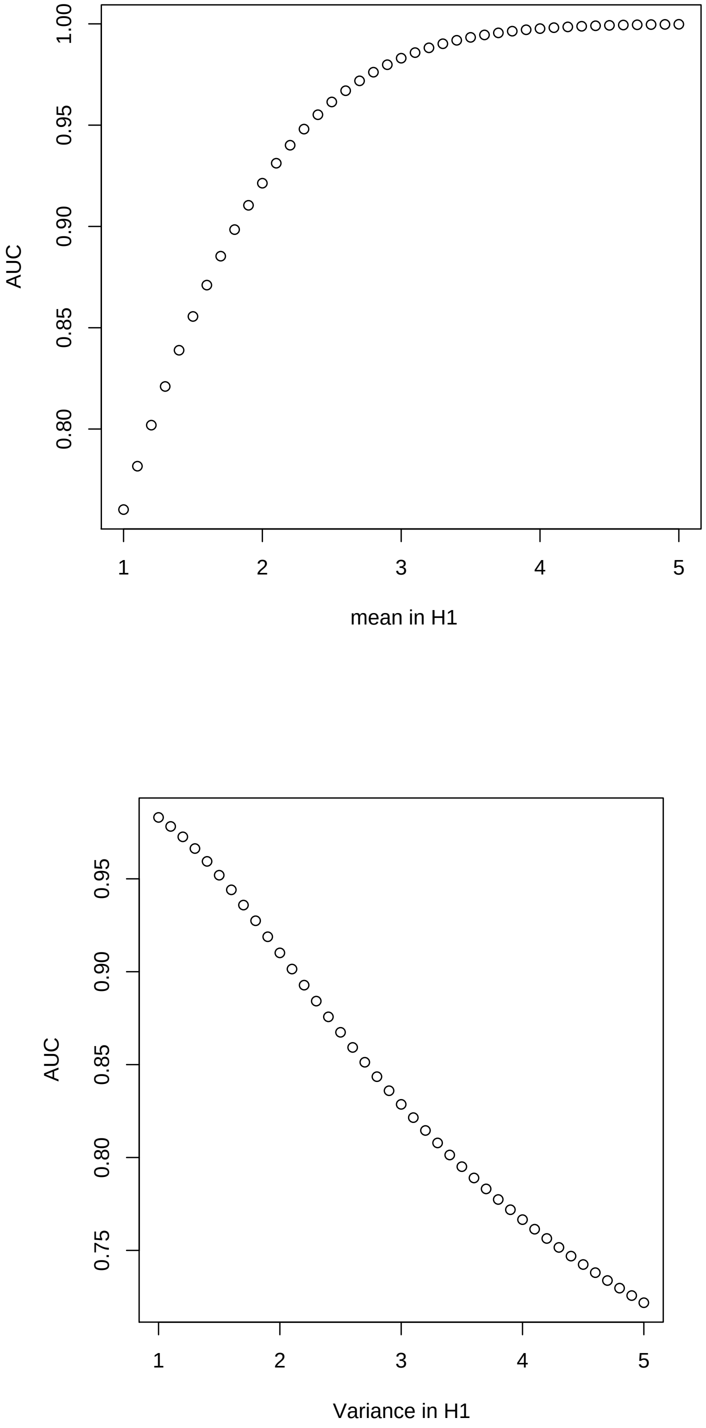


Figure 1.3: AUC vs. increasing mean or variance in the H_1

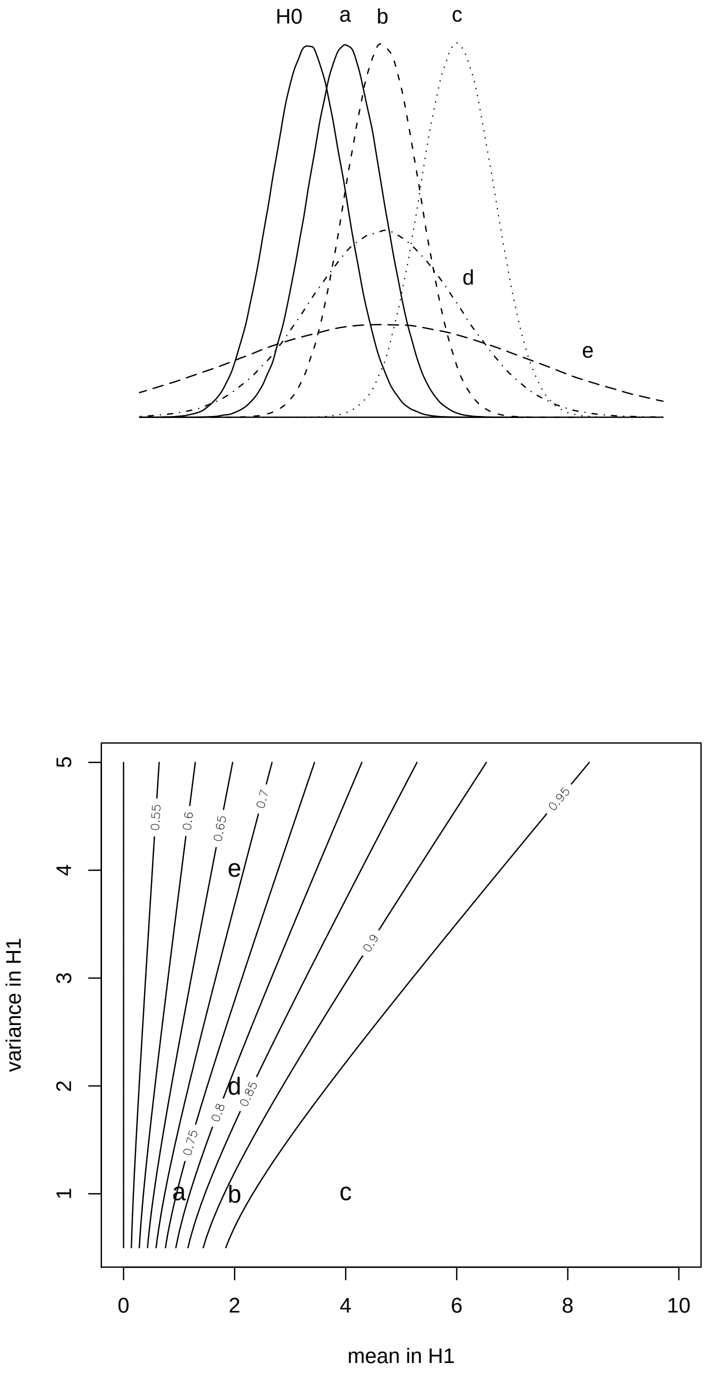


Figure 1.4: Contour plot of AUC as both mean and variance change in H_1.

sites. We also compare the GAM results and performance of the GAM methodology with those based on kernel density estimation of the early-to-late relative risk surface. Finally, we assess spatial performance of detection systems using decision fusion theory for the situation where a detection system can be comprised of a few expensive, precise detectors and many inexpensive, imprecise detectors in chapter six. In this thesis, we not only add new methods of evaluating cluster detection methods, but also bridge spatial performance of surveillance systems and sensor detection systems.

Chapter 2

Cluster Detection Methods

This chapter summarizes a set of statistical techniques for detecting spatial anomalies. We will use these to build spatial detection systems in subsequent chapters.

2.1 Detecting Clusters vs. Detecting Clustering

A cluster defines a collection of events that is different from the rest of a collection of events, while clustering defines a tendency for cases to gather together (25). Detecting clusters involves finding whether one or more collections of events are significantly different from the null hypothesis of no cluster, while detecting clustering tests whether cases tend to occur together. Detecting clustering usually provides a single test statistic and a single p-value for the entire data set, but detecting clusters could involve multiple testing with more than one p-value, one for each potential cluster.

Methods for detecting clusters and clustering can further be distinguished between general tests and focused tests (25). A general test investigates patterns within the entire study area, and a focused test is interested in a potential abnormal region, such as a neighborhood close to a waste site, or an area sharing the same water recourse. In this research, our primary research interest involves detecting clusters with general or focused tests.

There are many methods for identifying clusters in spatial epidemiological, statistical, and geographic analysis fields (8; 12; 25; 26; 28; 44; 46; 57; 58; 64). In the following sections we will introduce some popular methods used in public health surveillance. We will further evaluate the performance of these cluster detection methods in Chapter 4 and Chpater 5.

2.2 Spatial Scan Statistic

Joseph I. Naus studied scan statistics in detail in 1965 (34; 35). Since then, this method has been widely used in archaeology, epidemiology, geography, biology, ecologic, environment, sociology, and many other scientific and engineering fields (27). A scan statistic involves moving a "window" across the study area and comparing the incidence/prevalence rate observed within the window to that observed outside of the window. We test each window as a potential cluster. A spatial scan statistic was developed by Martin Kulldorff (45), and his popular SaTScan software system is free for users from the website http://wwww.satscan.org (45). The SaTscan software has been used in infectious diseases (38; 67), cancer (4; 47), diabetes (7), syndromic surveillance (29; 39), brain imaging (82), history (9), criminology (79), and many other fields.

In most applications, a circular window is used in the SaTScan under both Poisson and Bernoulli models of the counts within and outside of the window. The user is allowed to define the centers and radii of the circular windows under consideration, where each circle defines a potential cluster. Applications typically only consider circles covering less than half of the study area. Under the Poisson model, the null hypothesis assumes that the number of cases is proportional to the population size, that is, the rate is equal across the study region, while the alternative hypothesis assumes that the rate inside the window is higher than outside the window (45). The test statistic is the maximum value of the likelihood function associated with the potential cluster, which is proportional to:

$$\left(\frac{c_{in}}{E_{in}}\right)^{c_{in}} \left(\frac{c_{out}}{E_{out}}\right)^{c_{out}} I(c_{in} > E_{in})$$

where c_{in}, c_{out} are the observed number of cases inside and outside the potential cluster, E_{in}, E_{out} are the expected number of cases, $I()$ in a indicator function, $I()$ equal to one if $c_{in} > E_{in}$, zero otherwise.

Under the Bernoulli model, the null hypothesis assumes that $p = q$, where p is the probability that an event occurs inside the window, while q is the probability that the event occurs outside of the window. The alternative hypothesis assumes that $p > q$ (45). The likelihood is proportional to:

$$\left(\frac{c_{in}}{n_{in}}\right)^{c_{in}} \left(\frac{n_{in} - c_{in}}{n_{in}}\right)^{n_{in}-c_{in}} \left(\frac{c_{out}}{n_{out}}\right)^{c_{out}} \left(\frac{n_{out} - c_{out}}{n_{out}}\right)^{n_{out}-c_{out}} I()$$

where c_{in}, c_{out} are the observed number of cases inside and outside the potential cluster, n_{in}, n_{out} are the observed number of cases and controls inside and outside the potential cluster, $I()$ in a indicator function, $I()$ equal to one if $\frac{c_{in}}{n_{in}} > \frac{n_{in}}{n_{in}+n_{out}}$, zero otherwise.

The distribution of the maximum observed likelihood ratio statistic across all windows under the null hypothesis can be simulated by Monte Carlo sampling, providing a single p-value associated with the observed maximum, i.e. the most likely cluster.

A space-time scan statistic for surveillance is also implemented in the SaTScan software (49). Cylinders are used as the space-time scanning window in this model. The base of the cylinder represents space, and the height of the cylinder represents the length of time. The null hypothesis of a space-time permutation model assumes that there is no space-time interaction in disease incidence. Recently, the SaTScan was extended to also allow elliptical cylinders as well (48).

2.3 Upper Level Set Scan Statistic

The upper level set (ULS) scan statistic is an extension of the circle-based scan statistic (20; 21). Circles used in the SaTScan methods may have low power for detecting irregularly shaped clusters, or may include smaller or larger areas than necessary. In addition, the SaTScan approach may not be able to address clusters defined on the network, such as stream and highway systems, instead of geographical regions. On the other hand, Upper level set scan statistics can be used to detect arbitrarily shaped but still connected hotspots.

Similar to the SaTScan, the ULS computes the likelihood function associated the potential cluster defined by the connected cells.

$$L(Z) = \max L(Z, p_0, p_1) = L(Z, \hat{p_0}, \hat{p_1})$$

where Z is the potential cluster region, and inside Z all cells have the probability p_1 of experiencing an event. Outside of Z, all cells have the probability p_0 of experiencing an event, and $p_1 > p_0$.

It is usually very difficulty to do an exhaustive search for all potential hot spots in a large region. The primary difference between the spatial scan statistic and the upper level set scan approach lies in the set of potential clusters under consideration. The SaTScan approach uses expanding circles to reduce the number of potential hot spots, and the Upper level set scan uses connected components in the upper level sets to reduce the searching space. The connected components could be physical or arbitrarily defined.

An upper level set is determined by the empirical cell rates. The empirical cell rate could be calculated by $G_a = Y_a/A_a$, where G_a is the empirical rate, Y_a is the number of events in cell a, and A_a is the number at risk in cell a. To construct the ULS tree, first order the rates G_a, from top to down to construct the tree. If the area of interest is finite, so is the number of rates. An upper level set is the collection of cells whose rate G_a exceeds value g, where g could be any value from the set of empirical cell rates. Elements in the upper level set may not be connected, but a candidate for a hotspot should be a region with connected cells.

Therefore, a zone with certain connected cells in an upper level set is a potential hotspot. The null hypothesis is that the rates are same for all regions, while the alternative hypothesis is that a hotspot has higher rate than other regions. Data under the null hypothesis are easily simulated, and the p-values could be obtained from the resulting Monto Carlo null distribution.

2.4 Ratio of Kernel Estimators

The third analytic technique under consideration is due to Kelsall and Diggle who developed a non-parametric density ratio method based on kernel estimation (31; 32; 33). It is used for case and control point data. For a case/control study, we assume that the collections of case and control locations s follow heterogeneous Poisson processes with intensity functions $\lambda_1(s)$ and $\lambda_2(s)$, respectively. The log intensity ratio is defined by

$$\rho(s) = \log\{\lambda_1(s)/\lambda_2(s)\}.$$

When conditioning on the number of case and control sites n_1 and n_2, the data can be regarded as a pair of independent random samples with bivariate distributions across the study area with probability densities $f(s)$ and $g(s)$, where $f(s)$ and $g(s)$

are proportional to $\lambda_1(s)$ and $\lambda_2(s)$.

$$\begin{aligned}\rho(s) &= \log\frac{f(s)/\int_R \lambda_1(s)ds}{g(s)/\int_R \lambda_2(s)ds} \\ &= \log\{f(s)/g(s)\} - \log\frac{\int_R \lambda_1(s)ds}{\int_R \lambda_2(s)ds}\end{aligned}$$

Since the second part is just a constant across the region R, the spatial variation in the log intensity ratio is proportional to the log density ratio, ie.,

$$\hat{\rho}(s) \propto \log\{\hat{f}(s)/\hat{g}(s)\}$$

where $\hat{f}(s)$ and $\hat{g}(s)$ are kernel estimators of $f(s)$ and $g(s)$ respectively.

A tolerance interval covers a proportion with a stated confidence, similarly to a confidence interval covers a parameter with a stated confidence. In order to provide a measure of statistical significance, Kelsall and Diggle proposed a tolerance interval for $\hat{\rho}(s)$ defined by Monto Carlo simulation under the null hypothesis.

The computation can be performed in R using kernel estimation functions available in the KernSmooth package (15; 51). We use a standard bivariate normal density for our spatial 2-dimensional data.

2.5 Generalized Additive Models

We next consider extensions to the kernel smoothing approaching that allow local covariate effects. Hastie and Tibshirani introduced Generalized Additive Models (GAMs) in 1984 based on nonparametric regression or smoothing techniques (37; 76; 77). GAMs are extensions of generalized linear models. In generalized linear models, we have

$$g(\mu) = \alpha + \sum_{i=1}^{p} x_i\beta_i$$

where $\mu = E(Y)$, Y is distributed according to a member of the exponential family, or with known mean-variance relationship, $g(.)$ denotes a link function, α is the intercept, x_i is the ith independent variable, and β_i is the ith parameter coefficient.

In GAMs, some or all parametric terms in GLM could be replaced by smooth functions, such as LOESS or cubic splines, but the additive form is still kept.

$$g(\mu) = \alpha + \sum_{i=1}^{q} z_i\theta_i + f_1(x_1) + f_2(x_2, x_3) + \dots$$

where $\mu = E(Y)$, Y is distributed according to a member of the exponential family, or with known mean-variance relationship, $g(.)$ denotes a link function, α is the intercept, z_i is the ith independent variable, and θ_i is the ith parameter coefficient, the $f(.)$s are smooth functions, and x_1, x_2, x_3 are independent variables in the smooth terms.

Most GAM applications use one-dimensional smoothing functions (75; 80), and regression splines, such as cubic splines, are widely used to represent smooth functions. Cubic splines are one popular way to represent one-dimensional smoothing functions. A set of knots are placed among the data points, then a set of cubic polynomial functions are fitted to each section, those cubic spline are joined at those knots to make the whole spline continuous up to its second derivative. A cubic spline can be written as

$$\begin{aligned} f(x) =& \frac{(x_{j+1} - x)}{h_j}\beta_j + \frac{(x - x_j)}{h_j}\beta_{j+1} \\ &+ \frac{1}{6}[(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)]\delta_j + \frac{1}{6}[(x - x_j)^3/h_j - h_j(x - x_j)]\delta_{j+1} \\ &\text{if } x_j \le x \le x_{j+1}. \end{aligned}$$

where $x_1 \dots x_k$ are knots, $h_j = x_{j+1} - x_j$, $\beta_j = f(x_j)$, and $\delta_j = f''(x_j)$.

If we estimate coefficients in GAMs by likelihood maximization, this will often lead to a overfitted wiggly model. To control the model smoothness, a penalty term is introduced in our objective functions, therefore, we maximize a penalized likelihood for the model.

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_j \lambda_i \beta^T S_j \beta, \tag{2.1}$$

where λ is the smoothing parameter, and $\lambda \in [0, \infty)$, and $\frac{1}{2}\beta^T S_j \beta$ is the penalty for the jth smoothing term. When $\lambda = 0$, we obtain an un-penalized regression spline, and when $\lambda \to \infty$ we obtain a straight line estimate for f. Choosing the smoothing parameter λ is crucial to model fitting. If λ is too small, the data will be undersmoothed, if λ is too large, the data will be oversmoothed.

As a result, GAMs are often fitted through a penalized iteratively re-weighted least squares method given the penalized matrix, and the optimal fit is obtained when it converges. GAMs can also be fitted by backfitting (77), which iteratively smoothing partial residuals from the model.

Besides the choice of the smoothing parameter, another challenge in fitting GAMs is how to represent the smooth functions. The choice of knots in cubic splines is subjective, and the cubic smooth spine is limited to one predictor variable. In spatial applications, it makes more sense to use 2-dimensional smooth functions to define surfaces.

Thin plate splines are suitable to multiple predictor variables (74). Thin plate splines also have the attractive property of being equivalent to best linear unbiased spatial predictions obtained via kriging (19; 54; 55). A thin plate smoothing spline for estimating a 2-dimensional data $y_i = g(x_1, x_2) + \epsilon_i$ minimizes

$$\parallel y - f \parallel^2 + \lambda J_{22}(f)$$

where λ is the smoothing parameter, and the penalty function for smoothing two predictor variables is

$$J_{22} = \int\int \left[\left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 \right] dx_1 dx_2.$$

When GAMs are applied to point data, as in a case and control study, one common link function is the logit, i.e.,

$$g(\mu) = \log \frac{P[Y_i = 1]}{1 - (P[Y_i = 1])}.$$

Compared to the ratio of density based on kernel estimation from previous section,

$$\rho(s) = \log\{\lambda_1(s)/\lambda_2(s)\},$$

the linear predictors of GAMs are related to the ratio of density, since

$$\begin{aligned} g(\mu|\text{event at } s) &= \log \frac{P[Y_i = 1|\text{event at } s]}{1 - (P[Y_i = 1|\text{event at } s])} \\ &= \log \frac{P[Y_i \text{ is a case}|\text{event at } s]}{P[Y_i \text{ is a control}|\text{event at } s]} \end{aligned}$$

and the predicted ratio of density is

$$\begin{aligned} \rho(s) &= \log \frac{f(s)}{g(s)} \\ &= \log \frac{Pr[\text{case event at s}]}{Pr[\text{control event at s}]} \\ &\propto \log \frac{\lambda_1(s)}{\lambda_2(s)} \end{aligned}$$

Therefore, both ratio of kernel density estimation method and GAMs should reveal similar results for point process data even though kernel density estimation method incorporate 2D binned kernel density smoothing while GAMs apply thin plate regression smoothing. However, GAMs offer the opportunity to incorporate parametric covariate terms, while kernel estimates do not.

In spatial epidemiology, GAMs have been applied to predict lung cancer rate in five countries (52), model HIV incidence (24), describe the cancer mortality rate trend in EU (56). GAMs have been implemented in the "gam" and "mgcv" packages in R, providing a convenient computing platform for our studies.

Chapter 3

Data Sets

In this chapter, we introduce two data sets we will use to motivate our proposed methods for assessing statistical performance in spatial systems.

3.1 Severe Cardiac Defects in Santa Clara County, California

In 1981-1982, two wells in Santa Clara County, California were contaminated with organic solvents. As a consequence, public concern arose regarding a perceived cluster of children born with severe cardiac anomalies in seven census tracts served by these two wells (crude relative risk =2.2) during the same time period (71). In addition, the spontaneous abortion rate was also reported to be higher in this region (43). Here we focus on the cardiac anomalies, and the data include 259 census tracks containing 20,799 live births with 71 experiencing severe cardiac defects (16). The location of each case and control is defined by the centroid of the census tract of birth residence due to confidentiality restrictions. Census tracts are land areas defined by the U.S. Bureau of the Census, and they vary in size. A census tract typically contains about 4,000 residents. Therefore, census tracts are usually geographically smaller in cities, but much larger in rural areas. The centroid is the center of the population mass of a tract. This data set has been previously analyzed by Shaw et. al. (71) and used to compare cluster detection methods by Ding et.al. (16) and Waller et.al. (42)

Figure 3.1 and Figure 3.2 shows the locations of cases and live births in this data set. In Chapter 4, we apply both the SaTScan and the Upper Level Set Scan methods to this data set to detect clusters and assess spatial performance of these two methods.

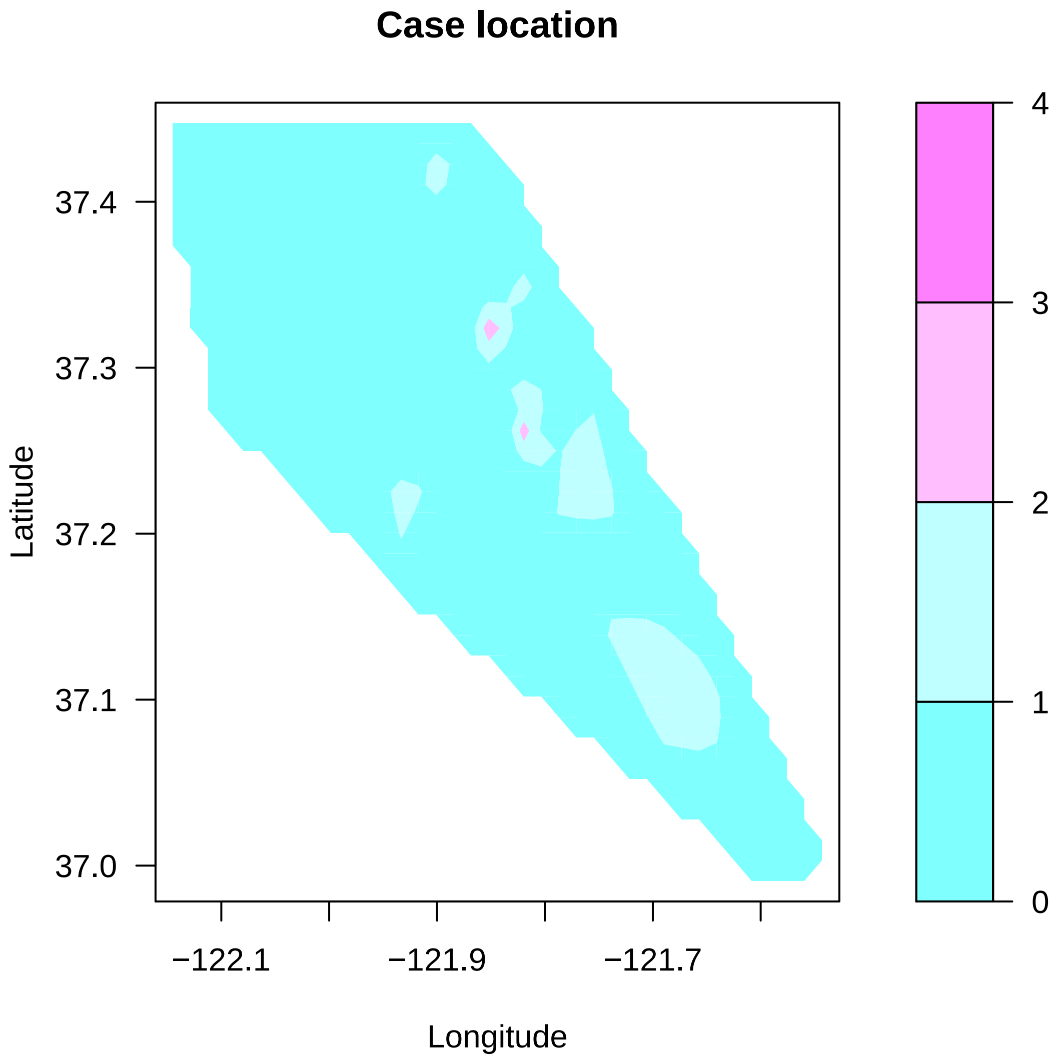


Figure 3.1: Contour map of locations of cases in Santa Clara County

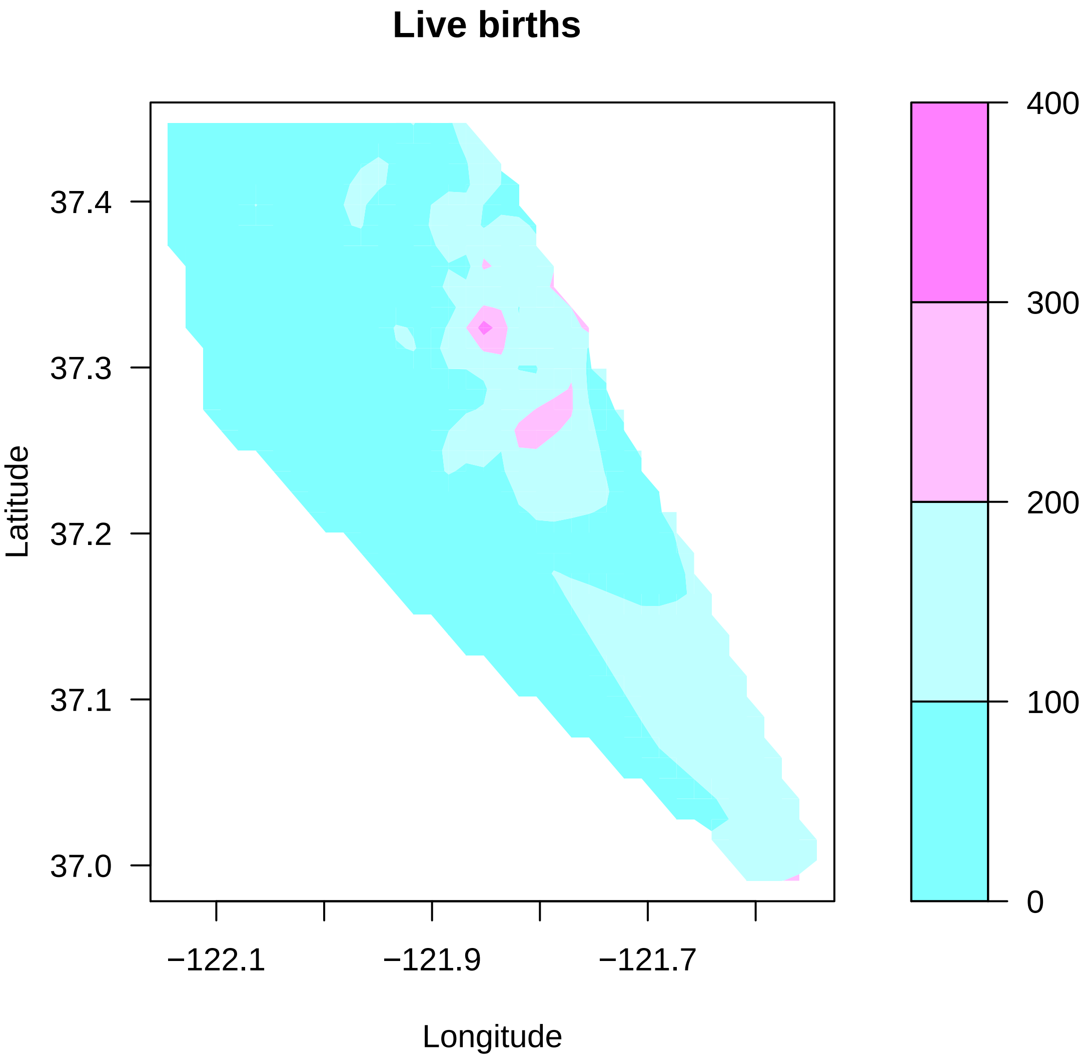


Figure 3.2: Contour map of locations of live birth in Santa Clara County

3.2 Anasazi Settlement sites on Black Mesa, Arizona

The Anasazi were an ancient population of Native Americans who lived in the southwestern United States including Arizona, Colorado, New Mexico and Utah (13; 36; 69). Even though this region is mainly composite with rugged mountain ranges, high plateaus, and a few rivers, the great ecological variety provided plenty of natural food to feed people for more than 3,000 years. The Anasazi were the largest and most well known among those lived in the area in ancient times. The modern day Pueblo tribes are descendents of the Anasazi. Our knowledge of the Anasazi primarily derives from archaeological remains, the records written by the Spanish explorers of the sixteenth century, or the traditions of the modern-day Pueblo, descendants of the Anasazi, since the Anasazi had no written history. Many archeological studies of past residential sites reveal a sharp growth in population up to 1000 A.D. followed by a sudden and fairly simultaneous abandonment of most sites around 1050 A.D. The question of what happened at that time remains a mystery.

There are thousands of Anasazi archaeological sites, and we are focusing on the sites studied by Peabody Coal Company's Black Mesa archaeological project. The Peabody Coal Company Black Mesa archaeological project is one of the largest archaeological projects in the American Southwest and was conducted in northeastern Arizona from 1967 to 1983. Hundreds of students and scholars worked on this project, and more than 700 settlement sites have been found and dated by pottery types and tree-ring dating. In the next chapter, we investigate whether the pattern of late sites dated between 950 and 1050 A.D., which spans the time from rapid population growth to the abandonment of settlements, are different from those early sites dated between 850 and 949 A.D. occuring prior to the sudden population growth. Our data include 100 early sites and 389 late sites. As the majority of the Anasazi in Black Mesa lived in small groups as farmers (13), we used geographic information systems to compute distance measures between sites and local streams and stream beds around those sites to explore the effects of these covariates on patterns of early and late sites.

The locations of early and late settlement sites appear in Figure **??**. In Chapter five, we apply ratio of density estimation, and GAM methods to this data set, then explore our spatial performance measures on them. The primary research questions of interest of this data set is comparison of the late and early site locations and the relationship between these patterns and proximity to rivers.

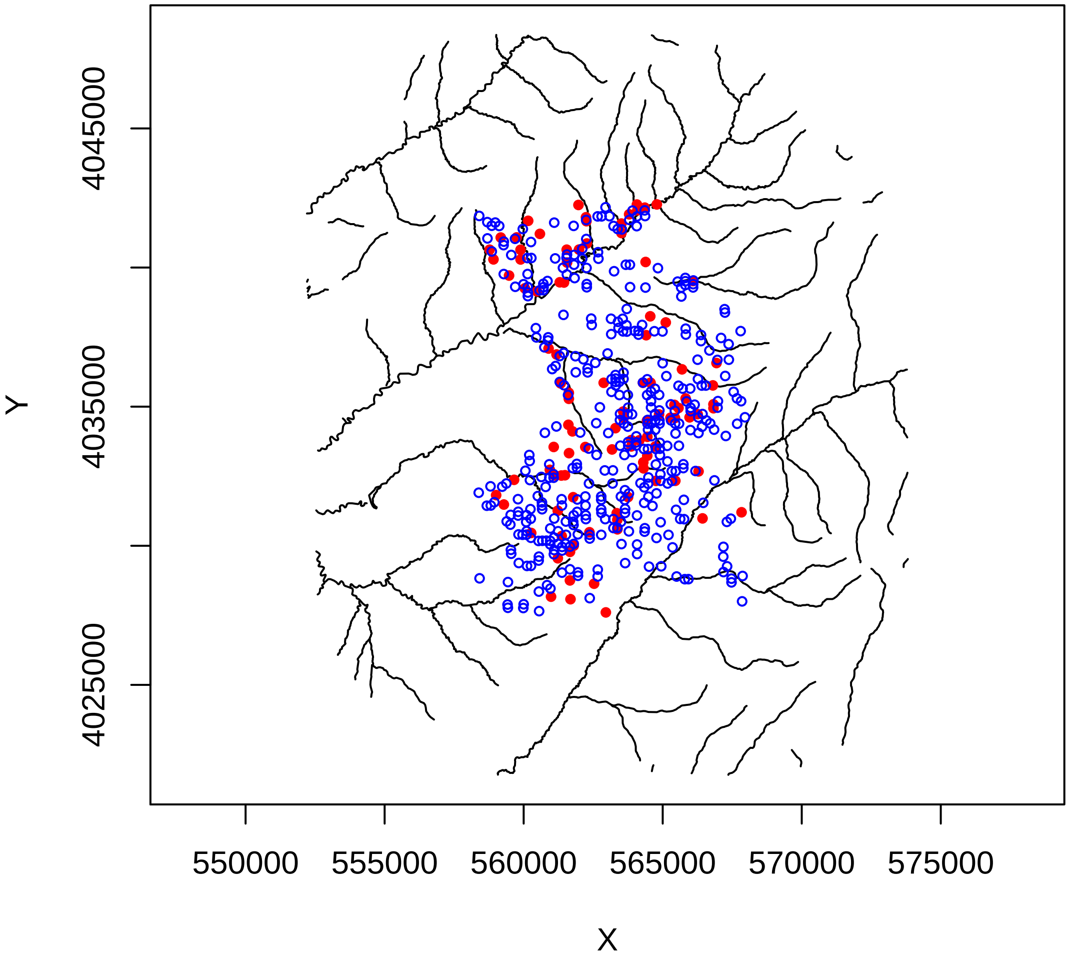


Figure 3.3: Settlement sites, red indicates early sites, and blue indicates later sites

Chapter 4

Spatial Measures of Performance: Regional Count Data

4.1 Introduction

In this chapter, we derive several spatial measures of statistical performance for cluster detection methods. Our motivating goal is to define how well statistical detection systems detect clusters occurring at various locations in data consisting of regional counts of disease. We introduce the spatially-referenced area under a receiver operating characteristic (ROC) curve (AUC) as a measure of local statistical performance summarizing local sensitivity and specificity. This work build on earlier examinations of spatially referenced power by Waller et. al (42).

The data set used in this section was introduced in the previous chapter. In 1981-1982, two wells in Santa Clara County, California were contaminated with organic solvents. Children born with severe cardiac anomalies in seven census tracts served by these two wells had crude relative risk of 2.2 compared to reference rates during the same time period (71). This data set includes 259 census tracts containing 20,799 live births with 71 experiencing severe cardiac defects (16). The location of each case and control is defined by the centroid of the census tract of birth residence due to confidentiality restrictions. Figure 4.1 and Figure 4.2 shows the locations of cases and live births in this data set.

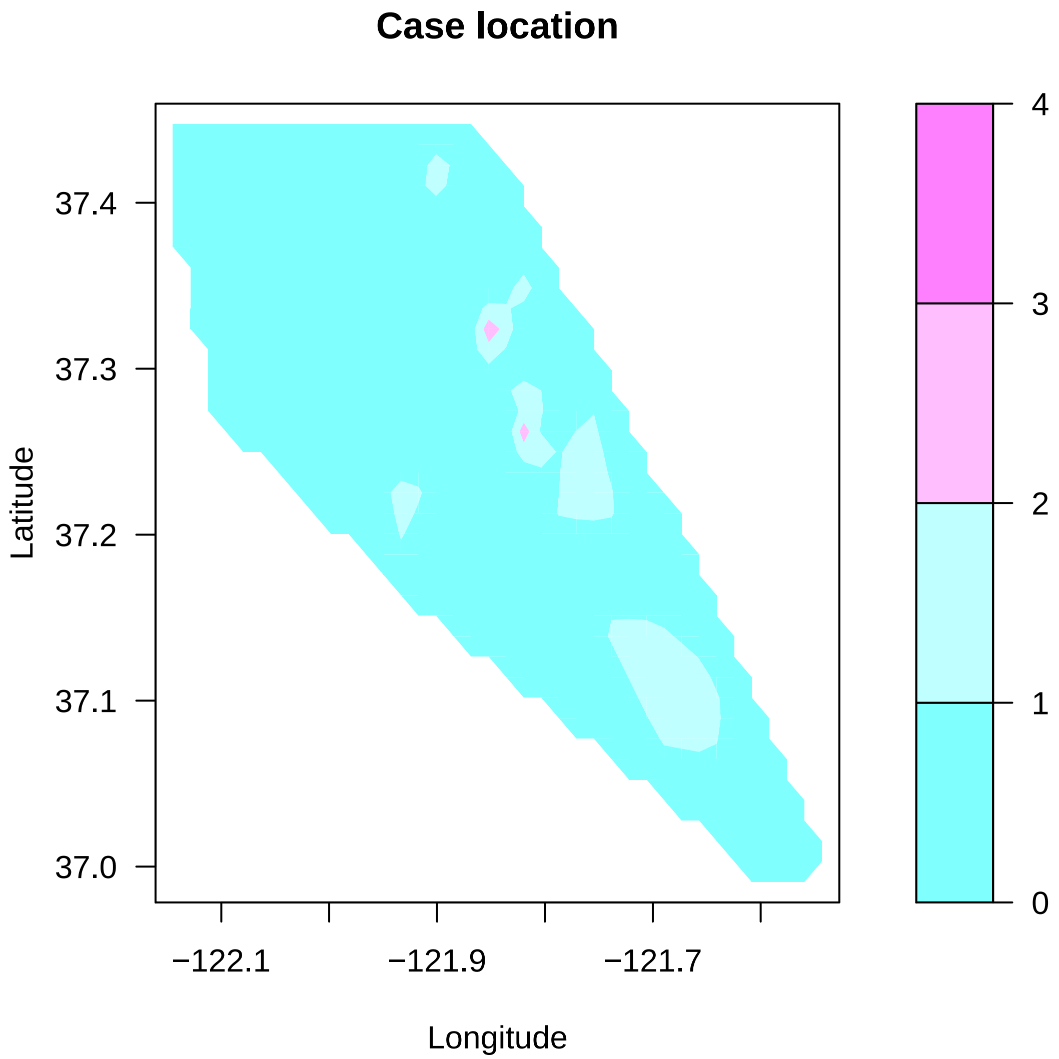


Figure 4.1: Contour map of locations of cases in Santa Clara County

We applied the spatial scan statistic implemented in the `SaTScan` software system (referred to as the SaTScan approach below), developed by Martin Kulldorff (45), and a second scan approach, the Upper Level Set Scan method, developed by G.P. Patil (20; 21), to the same simulated data sets based on the null and alternative

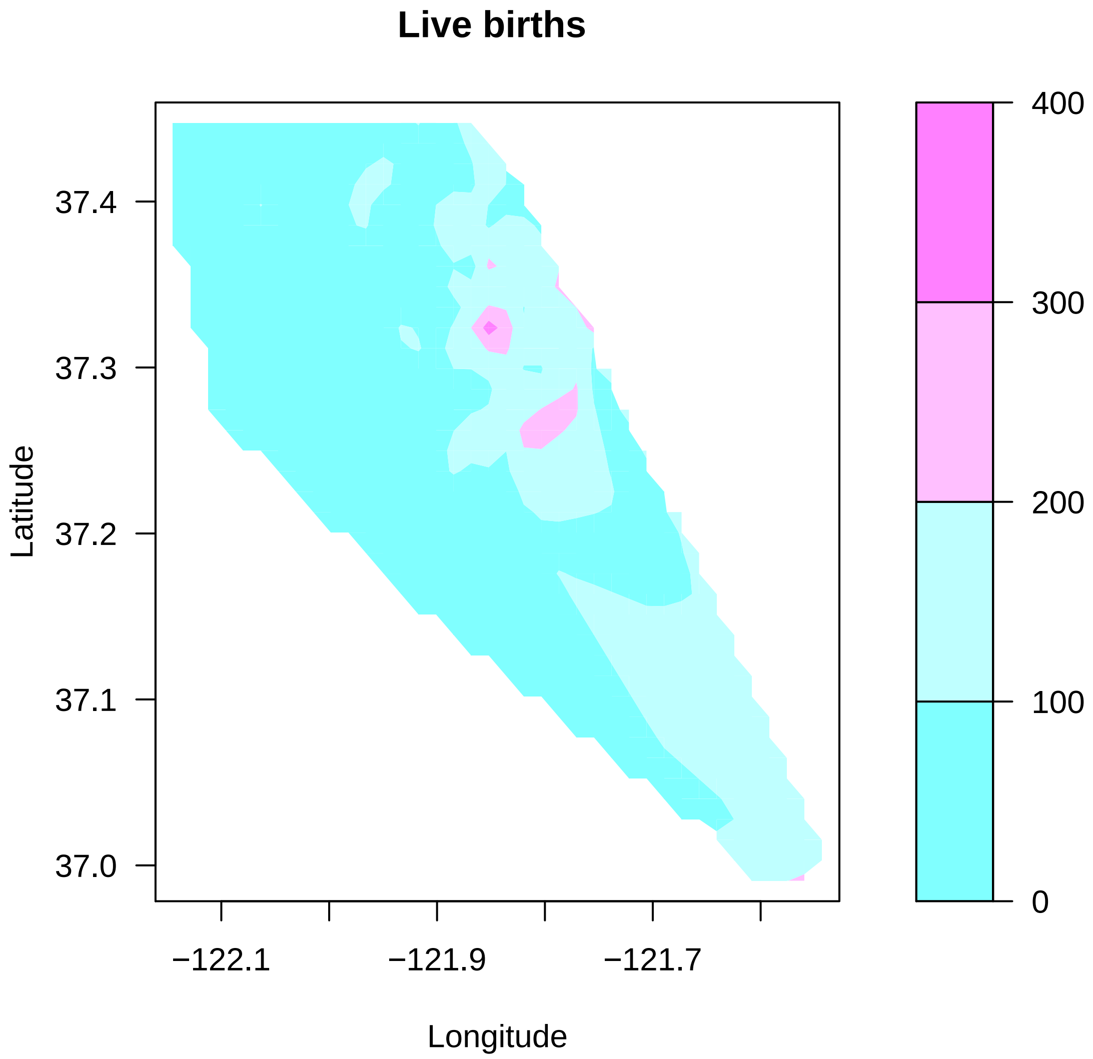


Figure 4.2: Contour map of locations of live birth in Santa Clara County

hypotheses to compute and plot AUC for each tract. The Upper Level Set (ULS) scan method requires information regarding connectivity between tracts. We use GeoDA, the software packages developed at the Spatial Analysis Lab, University of Illinois, to identify and output the connected tracts given a map with census tracts. The ULS method is executed by software written in the C programming language and provided by Dr Patil's research group. We modified the C code to fit our application here. We reviewed the results application of both SaTScan and ULS to this Santa Clara data, and compared spatial performance between them as well in the following sections.

4.2 Hypothesis

Under the null hypothesis, we assume the probability that a baby is born with a cardiac defect is the same everywhere. Therefore, the expected number of cases ($E(C_i)$) in each census tract is proportional to the number of live births (n_i) in the tract. Let γ be the baseline incidence rate, then, the null hypothesis could be represented as:

$$H_0 : C_i \text{ are independent Poisson random variables with } E(C_i) = \gamma n_i$$

Hill et. al. (42) consider an alternative hypothesis defined by

$$H_a : E(c_i) = \gamma n_i(1 + \delta_i \varepsilon)$$

where

$$\delta = \begin{cases} \text{1 if tract i is in the cluster} \\ \text{0 otherwise} \end{cases}$$

and $\varepsilon = RR - 1$, RR= relative risk of disease within the cluster.

For unknown γ, we redefine the hypotheses conditional on the sufficient statistic c_+:

$$\begin{cases} H_0 : C_1, \ldots, C_I | C_+ \sim \text{multinomial}(c_+, n_1/n_+, \ldots, n_I/n_+) \\ H_a : C_1, \ldots, C_I | C_+ \sim \text{multinomial}(c_+, \pi_1, \ldots, \pi_I) \end{cases}$$

where

$$\pi_i = \frac{n_i(1+\delta_i\varepsilon)}{\sum n_i(1+\delta_i\varepsilon)}$$

4.3 Test Statistics

4.3.1 Kulldorff's SaTScan Statistics

Kulldorff's spatial scan statistic (45) is the maximum of L_z, where L_z is the local likelihood ratio statistic. Based on a Poisson model,

$$L_z = \left(\frac{c_z}{\hat{\gamma} n_z}\right)^{c_z} \left(\frac{c_+ - c_z}{c_+ - \hat{\gamma} n_z}\right)^{c_+ - c_z} I[c_z > \hat{\gamma} n_z] \tag{4.1}$$

where $\hat{\gamma}$ is the estimated baseline incidence rate, i.e. $\hat{\gamma} = \frac{c_+}{n_+}$ and $I[c_z > \hat{\gamma} n_z]$ is an indicator function equal to 1 when the number of observed cases in zone z exceeds that expected under H_0, and is equal to zero otherwise. The most likely cluster is defined to be the zone z with the maximum L_z.

We apply the SaTScan method to the Santa Clara data, and the most likely cluster is a circular cluster centered at (-121.814, 37.1939) with radius 0.15 (Figure 4.3), including tracts listed in table 4.1. This cluster contains 7997 live births with 40 of them experiencing severe cardiac defects. Some tracts with zero event are included in this circular cluster. The log likelihood ratio is 4.636 with associated p-value of 0.562. Therefore, the most likely cluster is not significant.

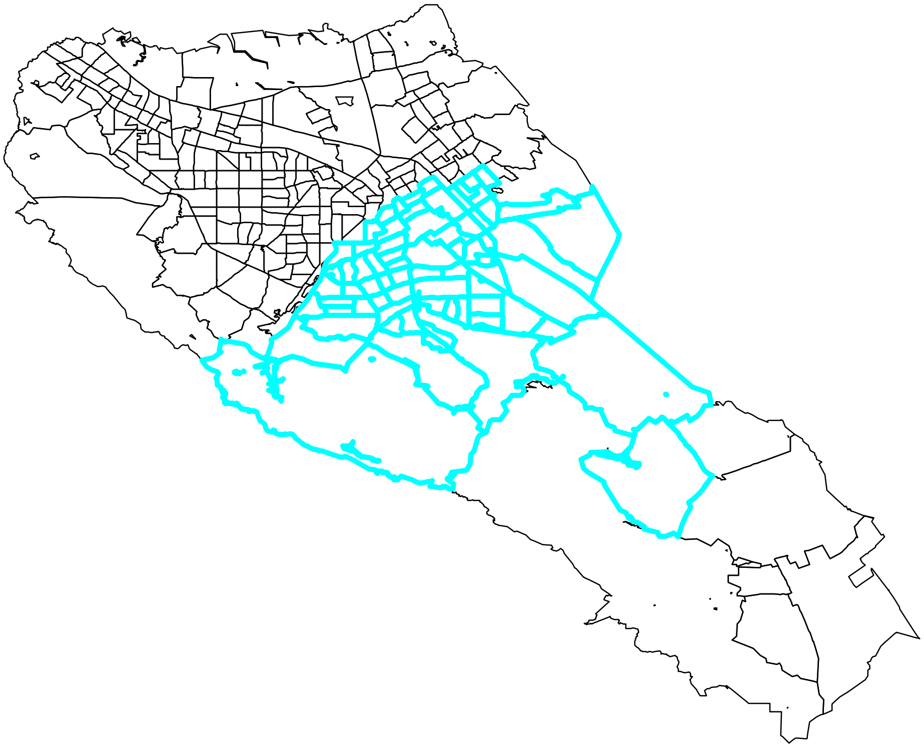


Figure 4.3: Most likely cluster identified by SaTScan in Santa Clara data.

4.3.2 Patil's Upper Level Sets Statistics

The Upper Level Set scan statistic (21) is the local likelihood ratio statistic, the same as the SaTScan. The difference between the Upper Level Set scan statistic and the SaTScan lies in the set of potential clusters considered by each approach. In the SaTScan, a set of circular windows with various radius are the potential zones; while in the Upper Level Sets Scan, a sets of connected regions are the potential zone to be detected. Therefore, even though the formulae of test statistics are the same, they could have very different identified zones.

We apply the ULS method to the Santa Clara data, and the most likely cluster is a cluster with 23 tracts, and none of these tracts has zero event (Figure 4.4). This cluster contains tracts listed in Table 4.2 This cluster contains 3,728 live births with 33 of them experiencing severe cardiac defects, and none of these tracts has zero events. The log likelihood ratio is 30.3935 with associated p-value of 0.40. Therefore, the most likely cluster is not significant.

Compared to the cluster identified by the SaTScan, the cluster identified by the ULS is smaller, and by construction it only contains non-zero event tracts. Three tracts (Figure 4.5) were in the ULS cluster but not in SaTScan.

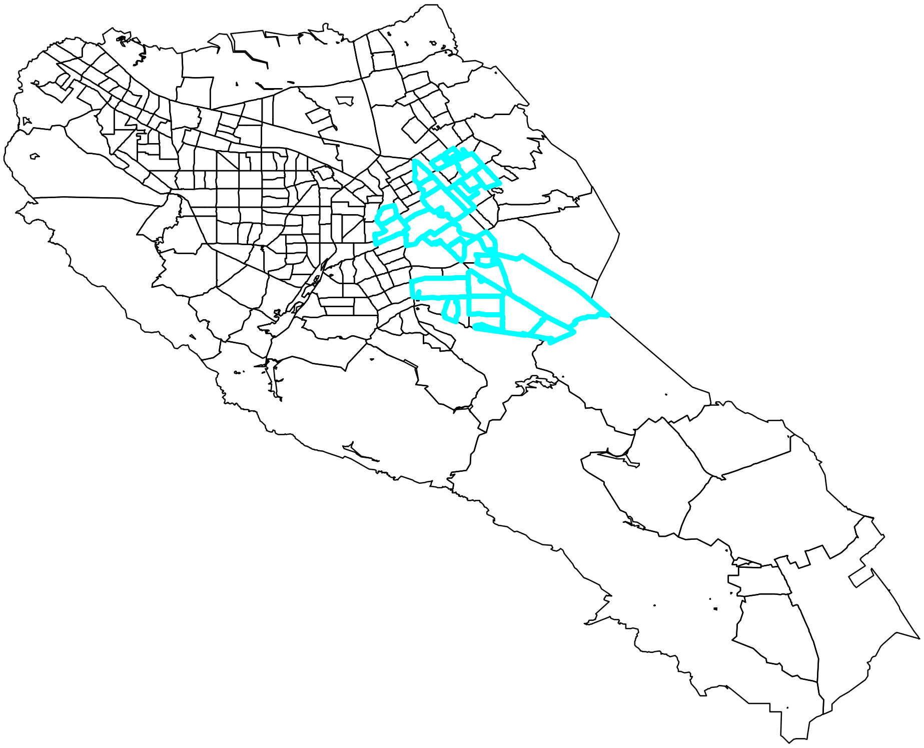


Figure 4.4: Most likely cluster identified by ULS in Santa Clara data.

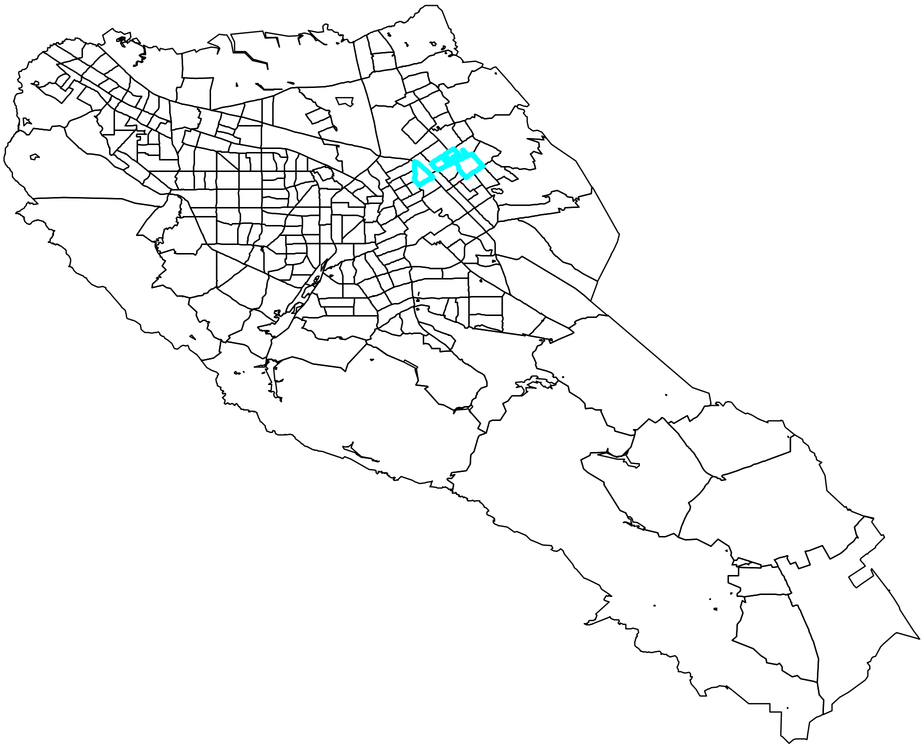


Figure 4.5: Tracts identified as most likely cluster by ULS but not by SaTScan in Santa Clara data.

Table 4.1: Cluster detected by SaTScan

Tract	Number of	Number	Tract	Number of	Number
	Live Birth	of Cases		Live Births	Cases
06085501500	195	1	06085501600	123	0
06085501700	117	0	06085501800	87	1
06085502300	64	0	06085502400	105	1
06085502500	88	0	06085502601	17	0
06085502602	57	0	06085502701	77	0
06085502702	95	1	06085502800	77	0
06085502901	69	0	06085502902	103	1
06085502903	48	0	06085502905	130	0
06085502906	66	0	06085502907	57	0
06085502908	100	1	06085503001	41	0
06085503002	33	0	06085503003	76	0
06085503101	168	0	06085503102	382	3
06085503103	120	1	06085503104	74	0
06085503203	144	1	06085503204	117	1
06085503205	162	0	06085503206	247	1
06085503207	65	1	06085503208	87	2
06085503304	129	0	06085503305	156	0
06085503306	73	0	06085503308	150	0
06085503309	222	0	06085503310	50	0
06085503400	269	2	06085503502	155	2
06085503504	150	1	06085503505	200	2
06085506801	51	0	06085506802	62	0
06085506803	93	0	06085506804	34	0
06085506900	80	2	06085507000	30	0
06085511800	54	0	06085511901	128	0
06085511903	44	0	06085511904	48	0
06085511905	25	0	06085511906	55	0
06085511907	52	1	06085511908	83	0
06085512002	290	1	06085512005	123	0
06085512006	102	0	06085512007	252	3
06085512008	180	1	06085512009	190	1
06085512010	123	0	06085512011	133	1
06085512012	132	2	06085512013	168	2
06085512014	92	0	06085512015	81	1
06085512016	125	0	06085512100	40	0
06085512301	182	2			
Total	7997	40			

Table 4.2: Cluster detected by ULS

Tract	Number of Live Births	Number of Cases
06085501400	134	1
06085501500	195	1
06085501800	87	1
06085502400	105	1
06085503102	382	3
06085503103	120	1
06085503203	144	1
06085503204	117	1
06085503207	65	1
06085503208	87	2
06085503400	269	2
06085503502	155	2
06085503504	150	1
06085503505	200	2
06085503705	177	1
06085504000	205	1
06085512007	252	3
06085512008	180	1
06085512009	190	1
06085512011	133	1
06085512012	132	2
06085512013	168	2
06085512015	81	1
Total	3728	33

4.4 Simulations

The simulation based on the null hypothesis is randomly assigning cases in the study region with the probability proportional to the number of live births in the tract. We simulated 9999 realizations.

For simulations under the alternative hypothesis, three types of clusters were defined as below.

4.4.1 Compacted Circular Cluster

The first type of cluster is a compacted circular cluster, and it is based on a tract and its six closest neighbors defined by the Haversine Formula, which is the distance on the globe:

$$\begin{aligned}
R &= \text{earths radius (mean radius} = 6,371\text{km)} \\
\Delta\text{lat} &= \text{lat}_2 - \text{lat}_1 \\
\Delta\text{long} &= \text{long}_2 - \text{long}_1 \\
a &= \left(sin(\frac{\Delta\text{lat}}{2}) \right)^2 + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \left(\sin \frac{\Delta\text{long}}{2} \right)^2 \\
c &= 2 \arcsin(min(1, \sqrt{a})) \\
d &= R \cdot c
\end{aligned}$$

Examples of this kind of circular cluster are shown in Figure 4.6 and Figure 4.7. The centriods of these seven cluster are inside a circle even though these tracts do not appear like a circle.

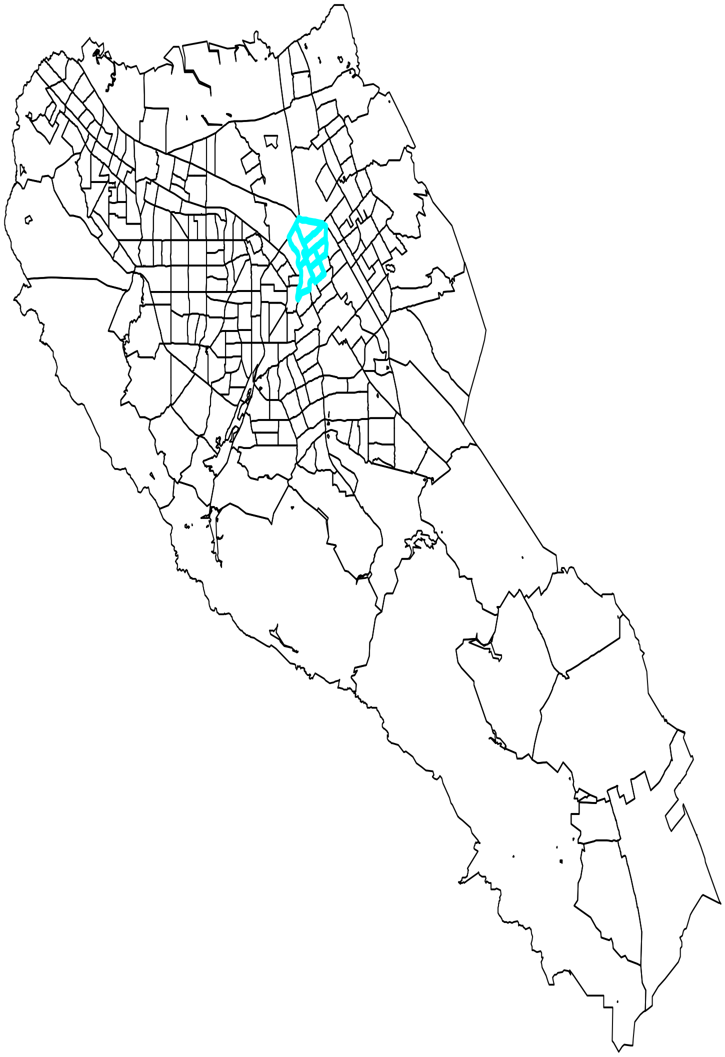


Figure 4.6: An example of circular cluster under the alternative hypothesis in the Santa Clara data.

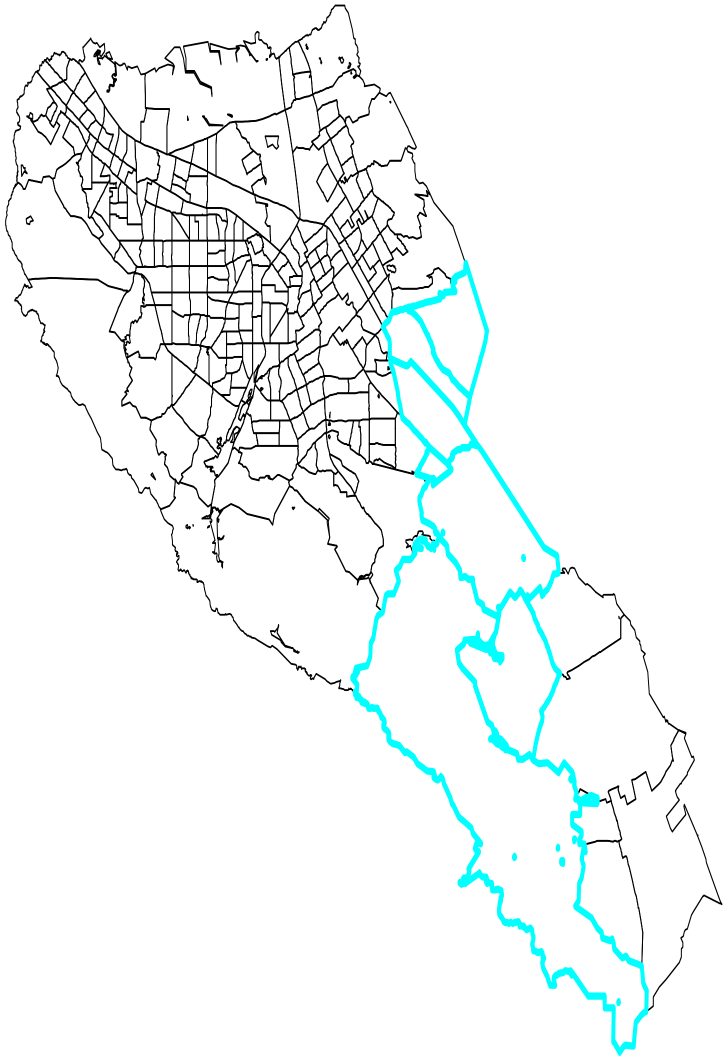
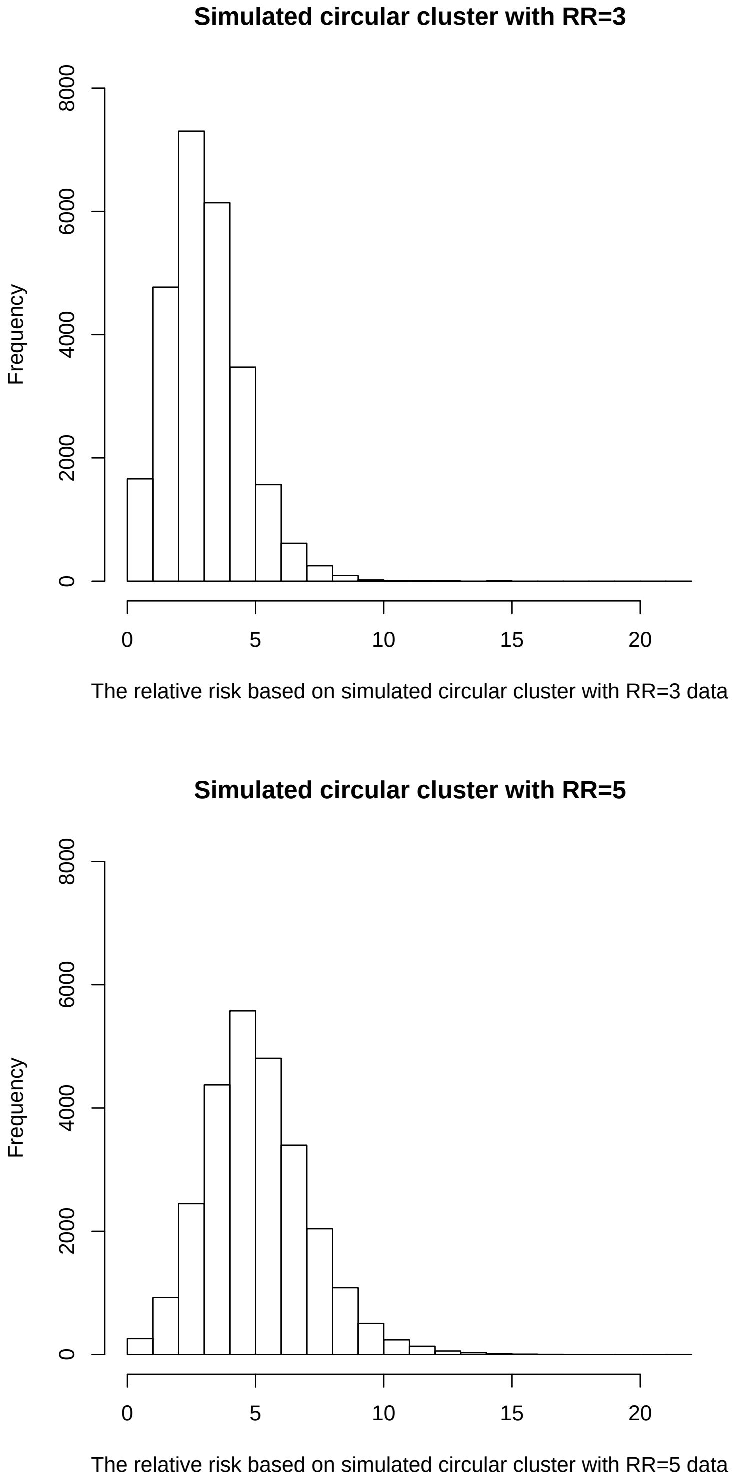


Figure 4.7: Another example of a circular cluster under the alternative hypothesis in the Santa Clara data.

We simulated two sets of data for this circular based cluster with relative risk at 3 and 5, respectively. The empirical histograms of relative risk from simulated data in Figure 4.8 show that the mean of the relative risk was 3.0 with range from 0 to 14.3, and 5.1 with range from 0 to 21.1, respectively. There are almost always tracts in the cluster experiencing zero events due to rareness of events as shown in Figure 4.9.



Simulated circular cluster with RR=3

Simulated circular cluster with RR=5

Figure 4.8: Histograms of the relative risk based on the simulated data of the alternative hypothesis in Santa Clara data. The top panel is based on simulated data with the relative risk of three, and the bottom panel is based on the simulated data with the relative risk of five.

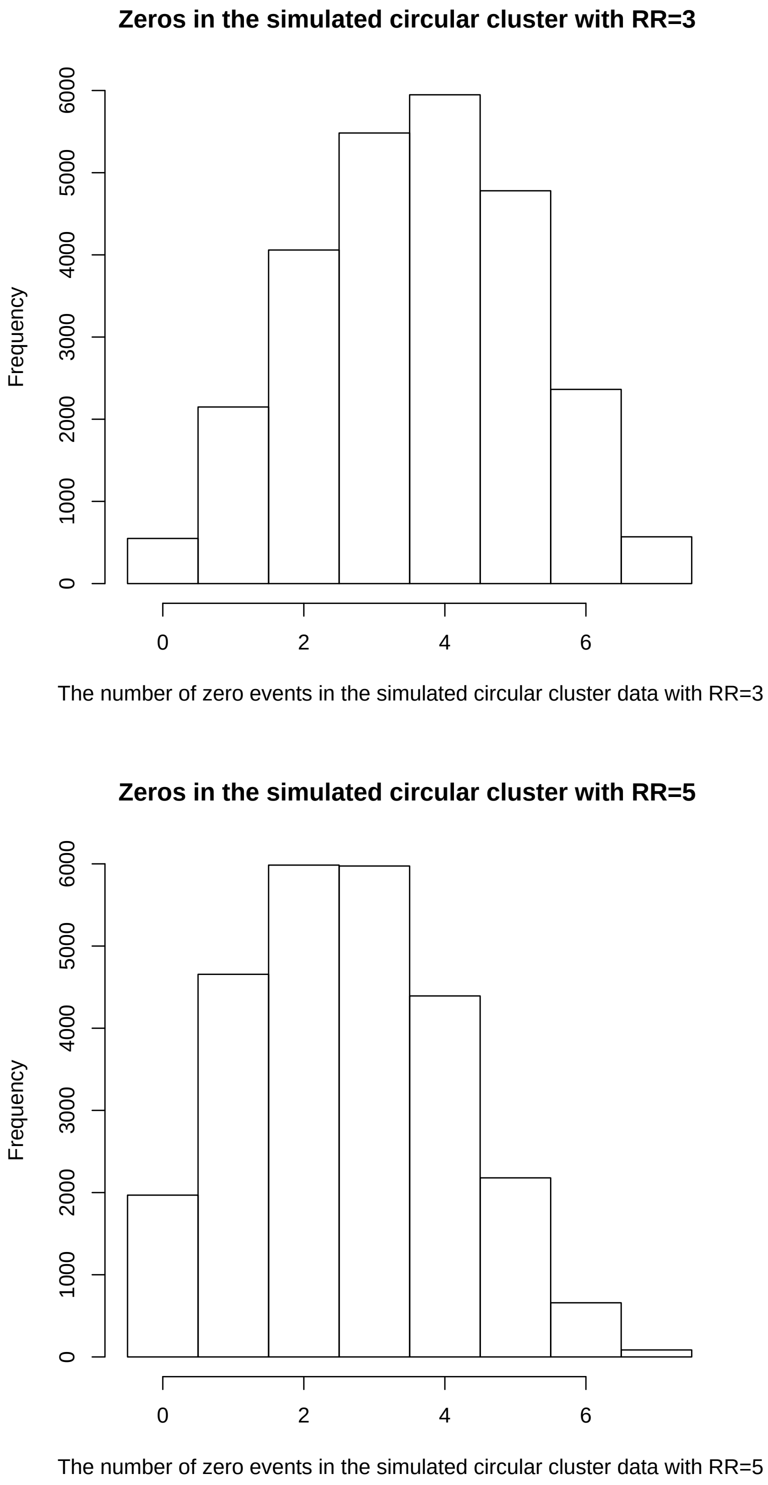


Figure 4.9: Histograms of the number of zero events based on the simulated data. The top panel is based on simulated data with the relative risk of three, and the bottom panel is based on the simulated data with the relative risk of five.

4.4.2 Closest Neighbor Cluster

The second type of cluster consisted of seven tracts with non-circular clusters defined by the following algorithm:

- define tract i as the starting point in the cluster.
- Randomly choose one connected neighbor of i as the 2nd tract in the cluster.
- Randomly choose one connected neighbor of the 2nd tract excluding those already in the clusters as the 3rd cluster.
- ...
- Randomly choose one connected neighbor of the 6th tract excluding those already in the clusters as the 7th cluster.

There are three tracts whose associated clusters have only six tracts as all the neighbors for the 6th tract are already in the cluster. Examples of these closely connected clusters are shown in Figure 4.10 and Figure 4.11.

Relative risk was set at $RR = 3$ for the simulation. The histograms of empirical relative risk in Figure 4.12 show that the mean of the relative risk was 3.05 with range from 0 to 12.72. Again, there are many tracts in the cluster with zero events due to the rareness of events as shown in Figure 4.12.

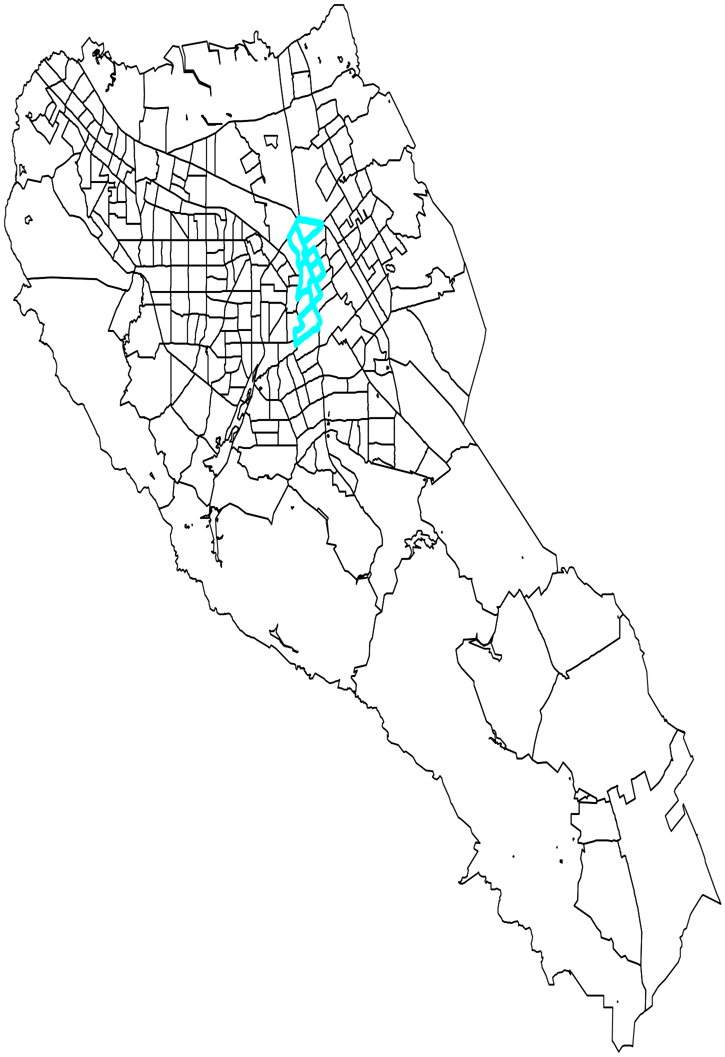


Figure 4.10: An example of a non-circular cluster under the alternative hypothesis in the Santa Clara data.

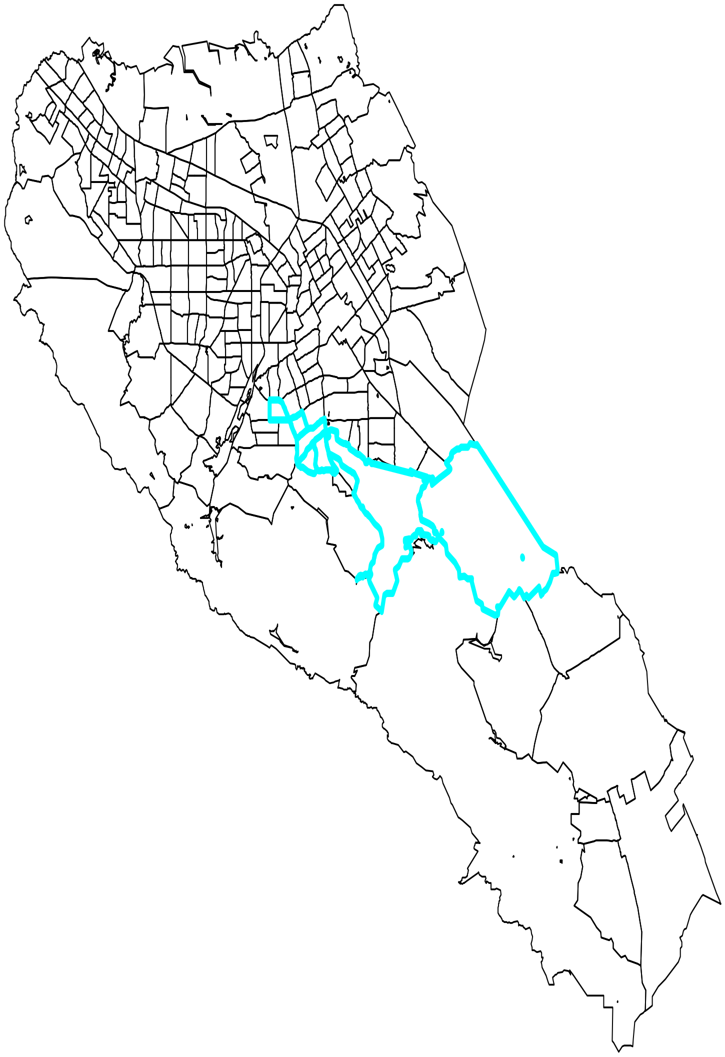


Figure 4.11: Another example of a non-circular cluster under the alternative hypothesis in the Santa Clara data.

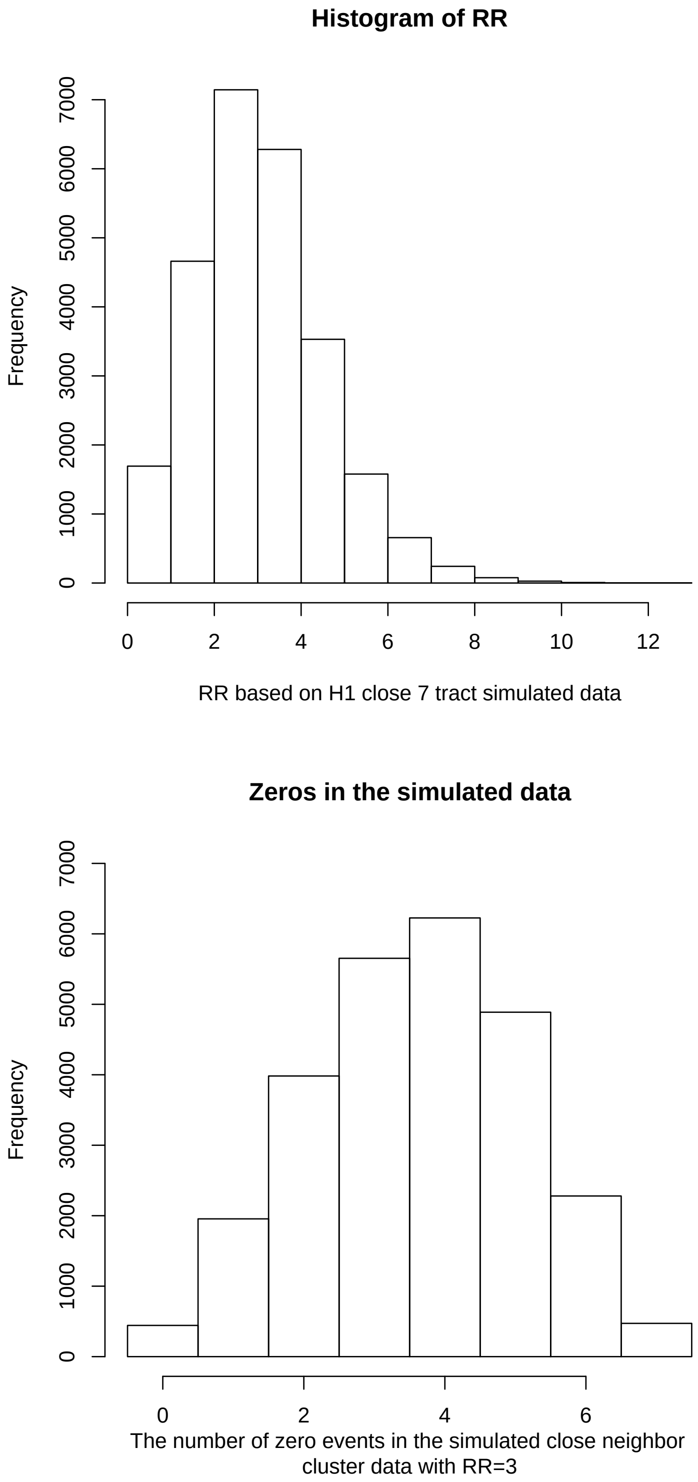


Figure 4.12: The top panel is the histogram of the relative risk estimated by simulation under the alternative hypothesis of closest connected neighbors in Santa Clara data with RR=3; the bottom panel is the number of zero events in the simulated closest neighbor cluster data.

4.4.3 Elongated neighbors cluster

The third type of cluster is more elongated than the previous one, and was defined by the following algorithm:

- define tract i as the starting point in the cluster.
- choose the farthest connected neighbor of i as the 2nd tract in the cluster.
- Choose the farthest connected neighbor of the 2nd tract excluding those already in the clusters as the 3rd tract.
- ...
- Choose the farthest connected neighbor of the 6th tract excluding those already in the clusters as the 7th tract.

There are two tracts whose associated clusters have less than seven tracts, one with five and the other with six tracts due to the last tract's neighbors already being in the cluster, and there are no other neighbors to include. Examples of these clusters are shown in Figure 4.13 and Figure 4.14.

Relative risk of the simulated data were set at $RR = 3$. The histograms of empirical relative risk in Figure 4.15 show that the mean of the relative risk was 3.05 with range from 0 to 12.72. Yet again, there are many tracts in the cluster with zero events due to the rarity of events.

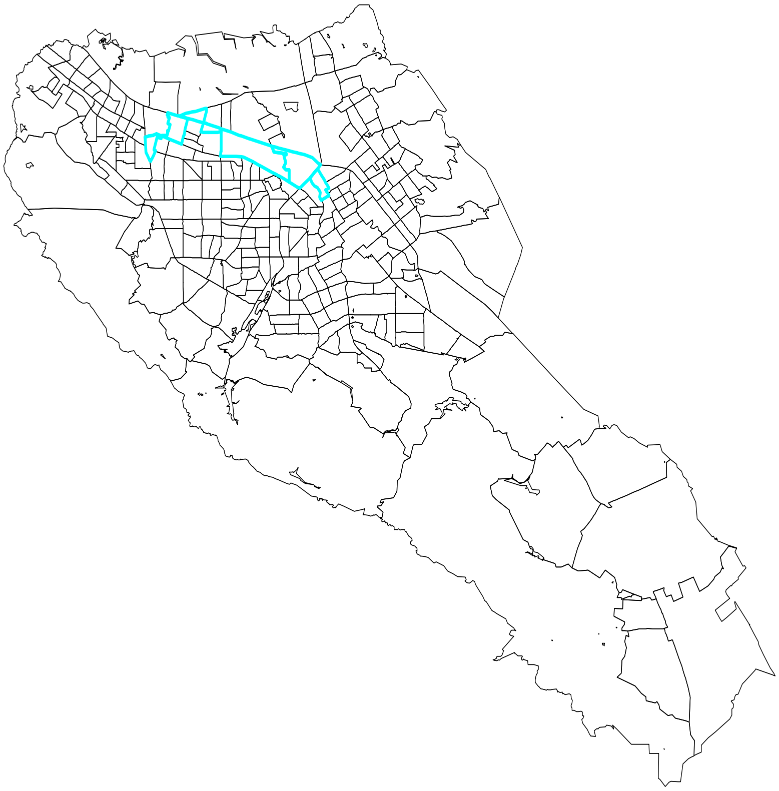


Figure 4.13: An example of a elongated cluster under the alternative hypothesis in the Santa Clara data.

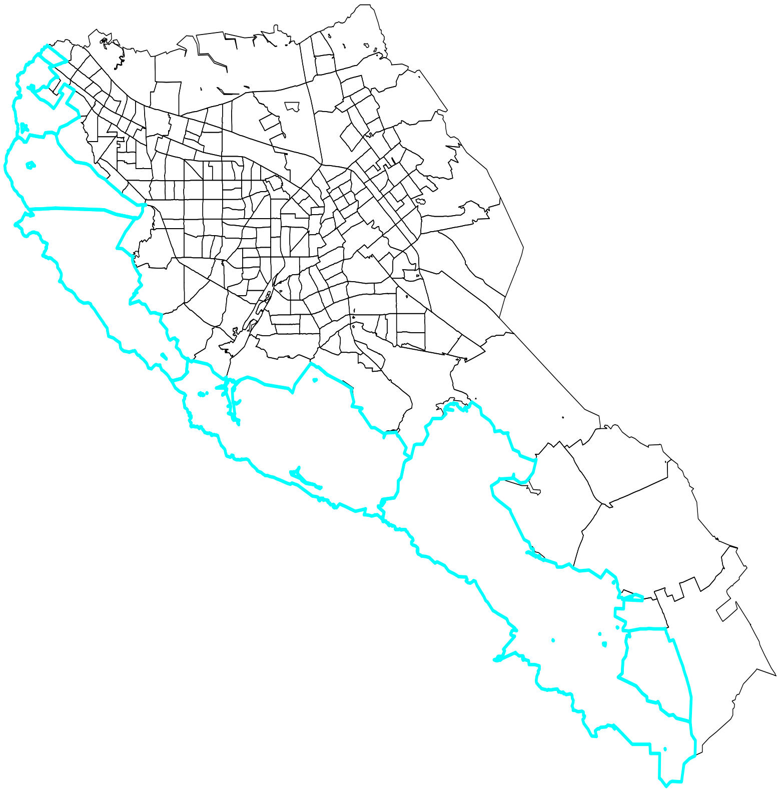


Figure 4.14: Another example of a elongated cluster under the alternative hypothesis in the Santa Clara data.

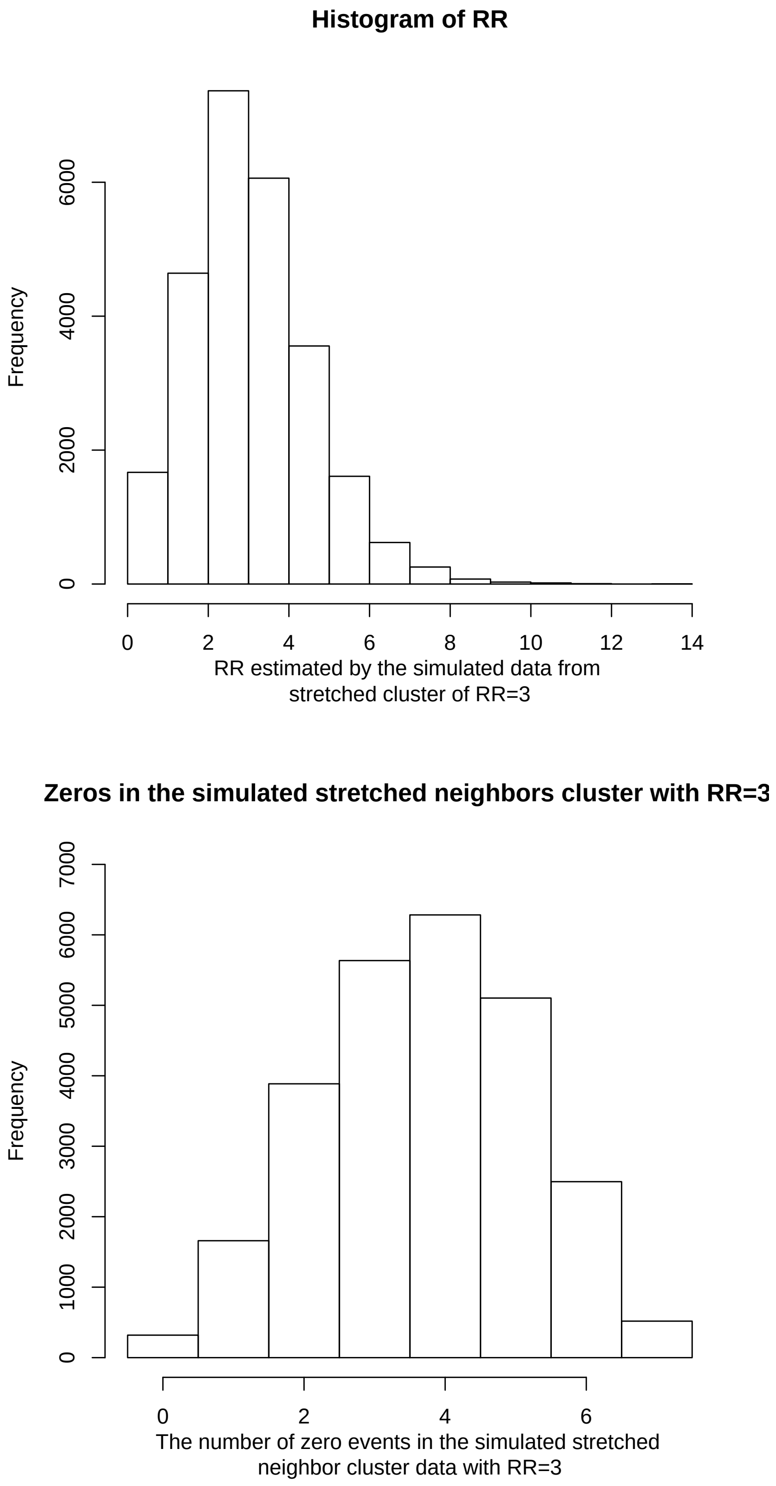


Figure 4.15: The top panel is the Histogram of the relative risk estimated by simulation under the alternative hypothesis of elongated neighbors in Santa Clara data with RR=3; the bottom panel is the number of zero events in the simulated elongated neighbor cluster data.

4.5 Defining the Spatial AUC for Regional Data

As a measure of performance, we consider a ROC curve associated with each tract's cluster. We estimate area under receiving operating characteristic curve via Monte Carlo simulations as detailed in the following sections.

4.5.1 Cut Points

First, we define cut points (critical values) for Kulldorff's scan statistic L_z associated with cumulative probabilities of 0.05, 0.1, 0.15, ... , 0.90, 0.95 based on the null distribution defined via Monte Carlo simulation as in Table 4.3.

Table 4.3: Cut Point defined by SaTScan from 9,999 simulated data based on the null hypothesis

probability	cut point
0.05	7.652
0.10	6.875
0.15	6.394
0.20	6.070
0.25	5.783
0.30	5.551
0.35	5.329
0.40	5.133
0.45	4.964
0.50	4.806
0.55	4.657
0.60	4.498
0.65	4.333
0.70	4.194
0.75	4.050
0.80	3.885
0.85	3.693
0.90	3.484
0.95	3.192

Next, cut points for the ULS statistics are based on the 9,999 simulations of the null hypothesis, and listed in table 4.4

Table 4.4: Cut Points defined by ULS from 9,999 simulated data based on the null hypotheis

probability	cut point
0.05	15.97
0.10	18.26
0.15	20.04
0.20	21.54
0.25	22.90
0.30	24.19
0.35	25.45
0.40	26.66
0.45	27.90
0.50	29.13
0.55	30.43
0.60	31.77
0.65	33.23
0.70	34.77
0.75	36.46
0.80	38.43
0.85	40.84
0.90	43.94
0.95	48.89

4.5.2 Estimating spatial AUC

For each tract's cluster,we estimate AUC based on these simulated data. The sensitivity is calculated as the proportion of the test statistics L_{max} values from simulated data sets exceeding each cut point based on the methods used. The area under the curve (AUC) is calculated via the Trapezoid Rule using Area $= \sum_{i=j}^{n} a_j = \sum_{j=1}^{n} \frac{h_{j+1}-h_j}{2}(b_j + b_{j+1})$, where h_j is the false alarm rate at cut point j, and b_j is the sensitivities at the cut point j.

The algorithm can be summarized as the following:

- Simulate 9999 data sets under the null hypothesis.

- Compute Kulldorff's or Patil's scan statistics L_{max} based on each simulated data set under the nulll hypothesis.
- Order L_{max}, and get cut points at 5th, 10th, ..., 95th percentiles.
- Simulate 100 data sets under the alternative hypothesis assuming that tract i is the center of the cluster.
- Compute Kulldorff's or Patil's scan statistics L_{max} based on each simulated data set.
- Calculate the proportion that L_{max} exceeding the cut points
- Compute AUC for tract i.
- Repeat for each tract.

In the following section, we presented spatial AUC for those three types of clusters defined previously.

4.5.3 Case I: compact circular cluster

As noted in section 4.4.1, two sets of data were simulated under the alternative hypothesis that there is a circular cluster consisting of one tract and its six closest neighbors defined by Haversine distance on the globe. Based on these simulated data, we calculated local AUC values by the algorithm stated above. Figure 4.16 shows the contour plot based on our local AUC values for the circular cluster at $RR = 3$. We note that the AUC varies with location, ranging from near-noninformative levels of 0.5 up to above 0.8. The map reveals areas where SaTScan would have better or worse ability to detect the type of clusters under consideration.

Figure 4.17 shows the contour plot based on our local AUC values based on the ULS on simulated circular cluster with RR=3. We note that AUC varies from 0.3 to

0.7. Usually, when AUC is below 0.5, people would flip the sign of the hypothesis, that is, instead of finding a region with higher risk, we may declare that a region with unexpected lower risk. Here we keep the original testing hypothesis of looking for higher risk regions. Therefore for those regions with AUC less than 0.5, there is little chance of detecting them even though they truly experience a higher risk.

In order to assess whether the observed variation in AUC is driven entirely by local variation in the number of live births, Figure 4.18 reveals a general trend of increasing AUC with increasing number of live births for both the SaTScan and the ULS, but also reveals substantial variation about this trend, indicating influences outside of simple variations in the local number of live births.

The seven tracts falling outside of the general point cloud are located at the far south end of the region as shown in map 4.19. In most locations the SaTScan outperforms the ULS as shown in Figure 4.20. We further examined the empirical mean number of zero events in the predefined seven tract cluster. Figure 4.21 shows the number of zero event affect performance for both SaTScan and ULS dramatically.

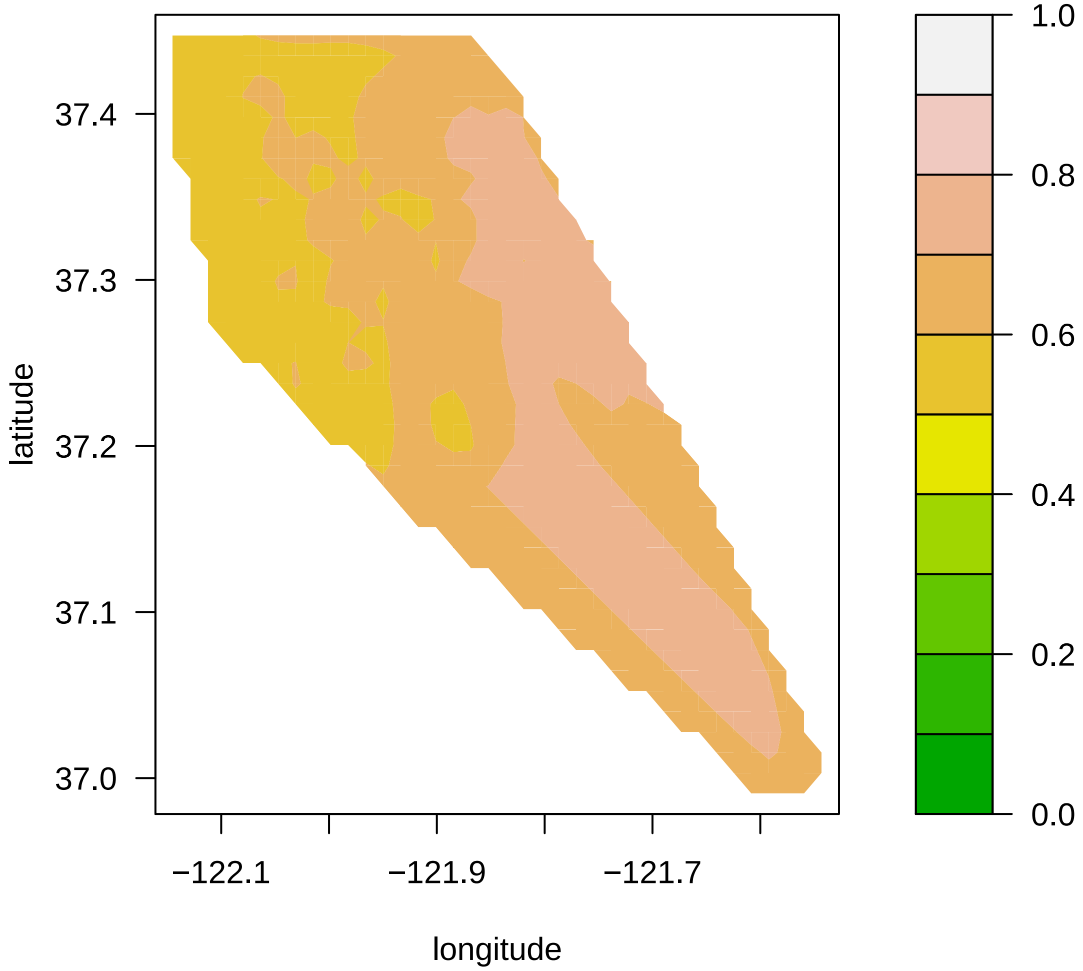


Figure 4.16: Contour map of AUC for SaTScan with RR=3 and simulated circular clusters.

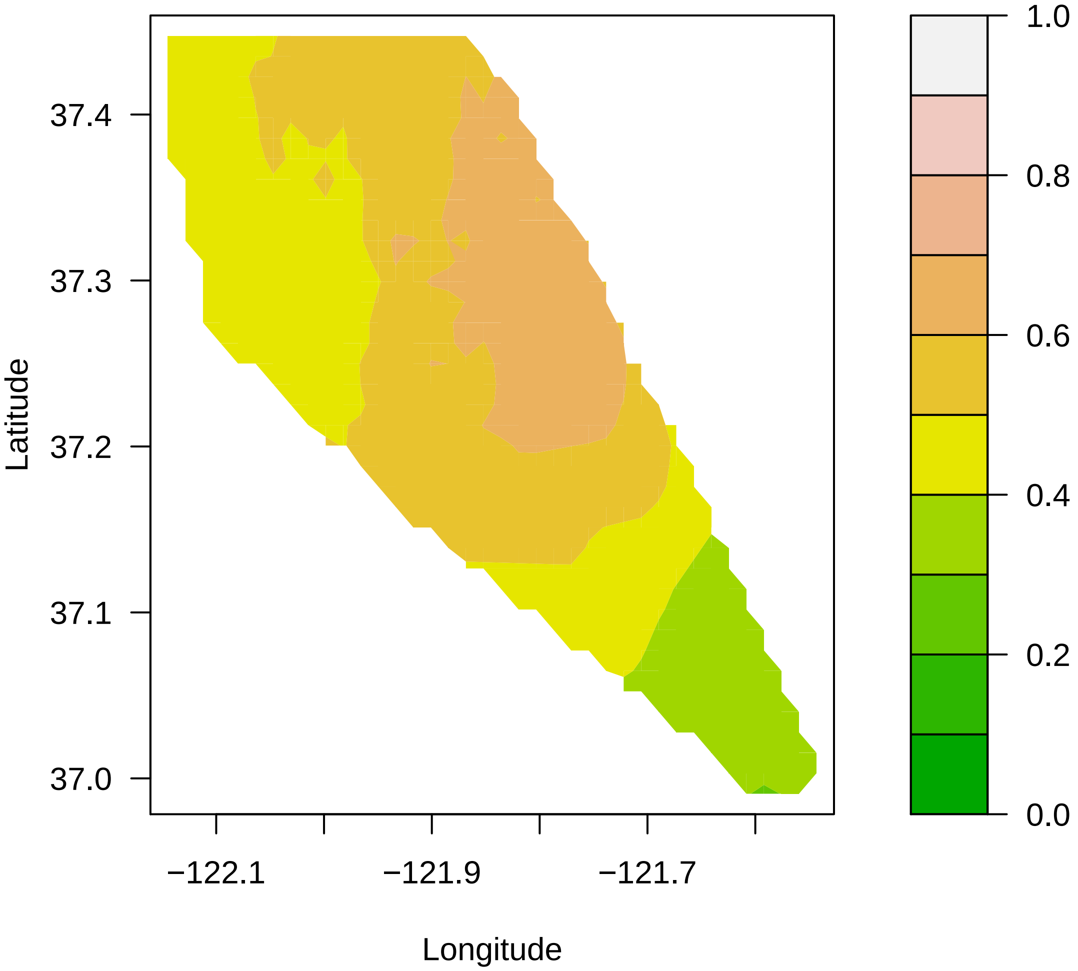


Figure 4.17: Contour map of AUC for ULS with RR=3 and simulated circular clusters.

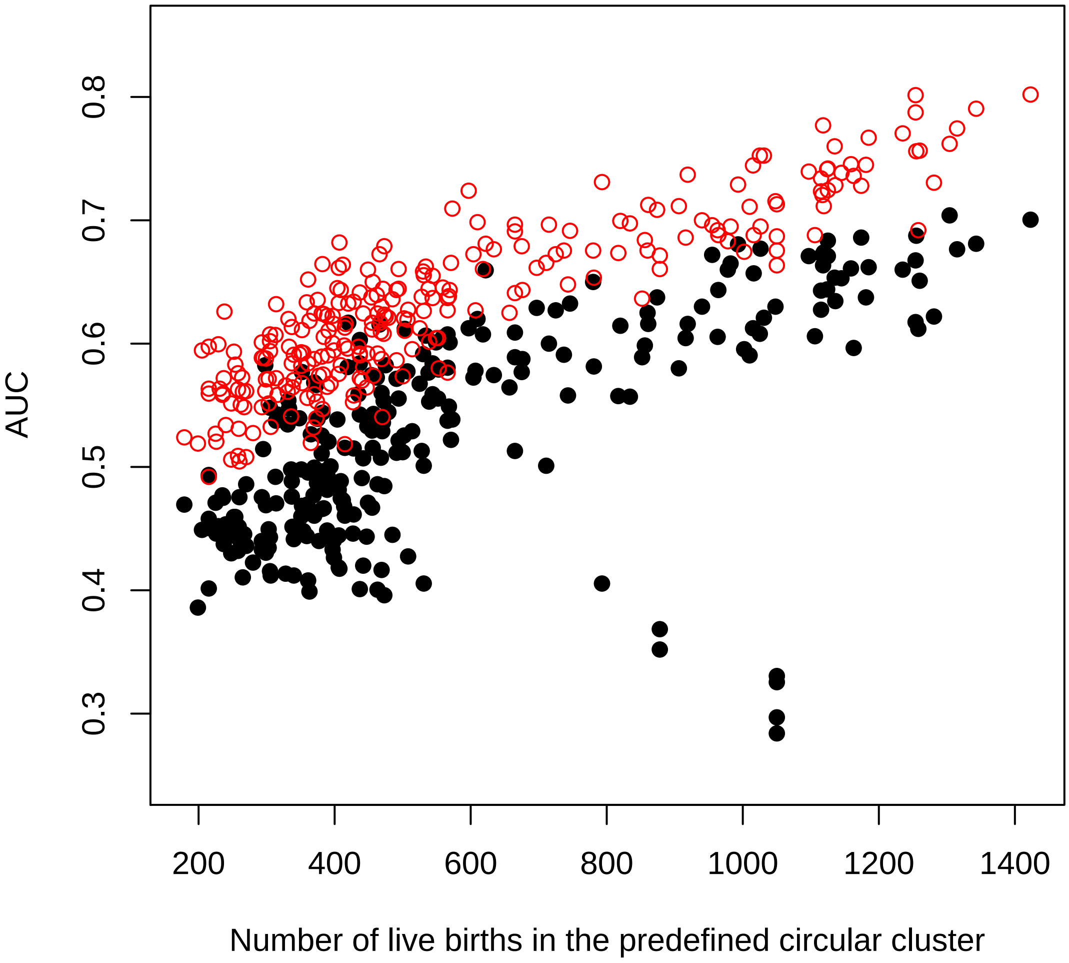


Figure 4.18: Scatter plot of AUC vs. the number of live births in the circular cluster. Red denotes AUC of SaTScan, and black denotes AUC of ULS with RR=3.

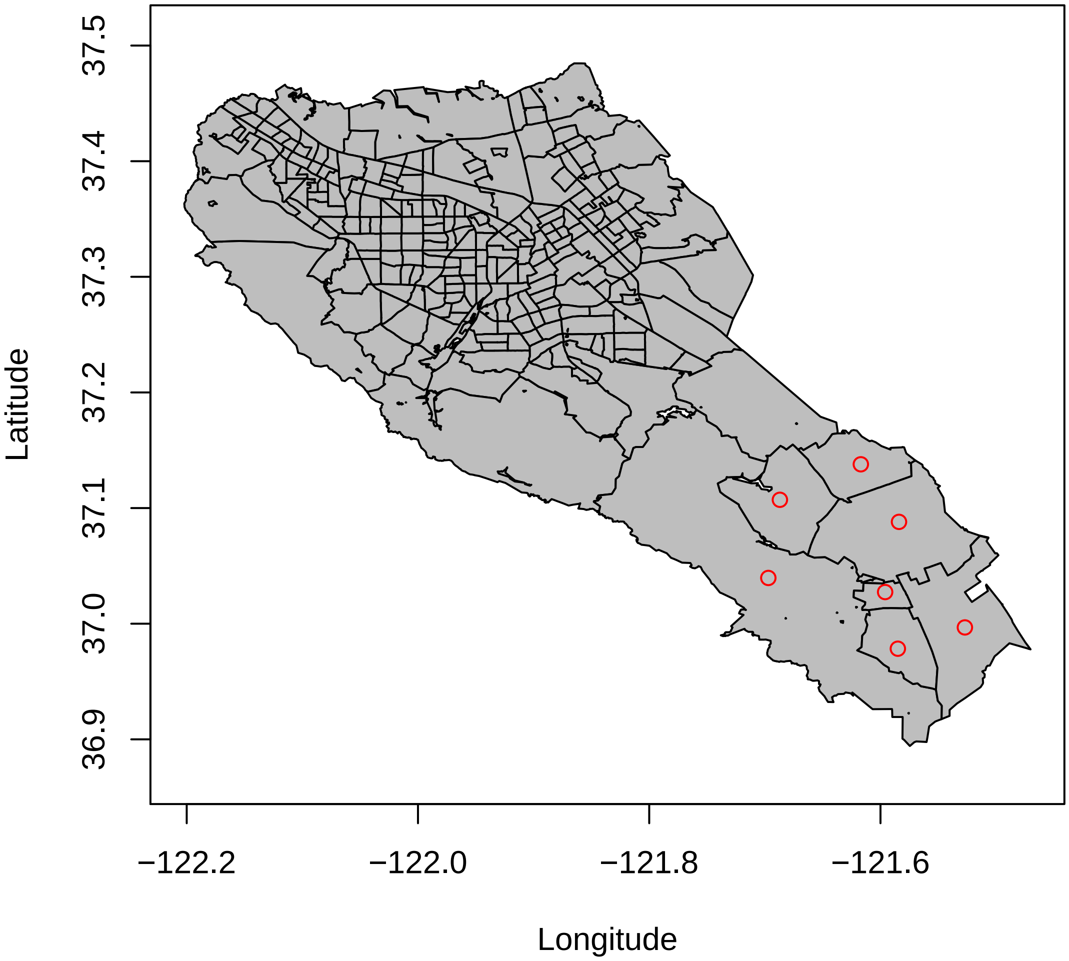


Figure 4.19: Map showing the seven tracts with different pattern of trending plot of AUC for ULS vs. the number of live births in the cluster.

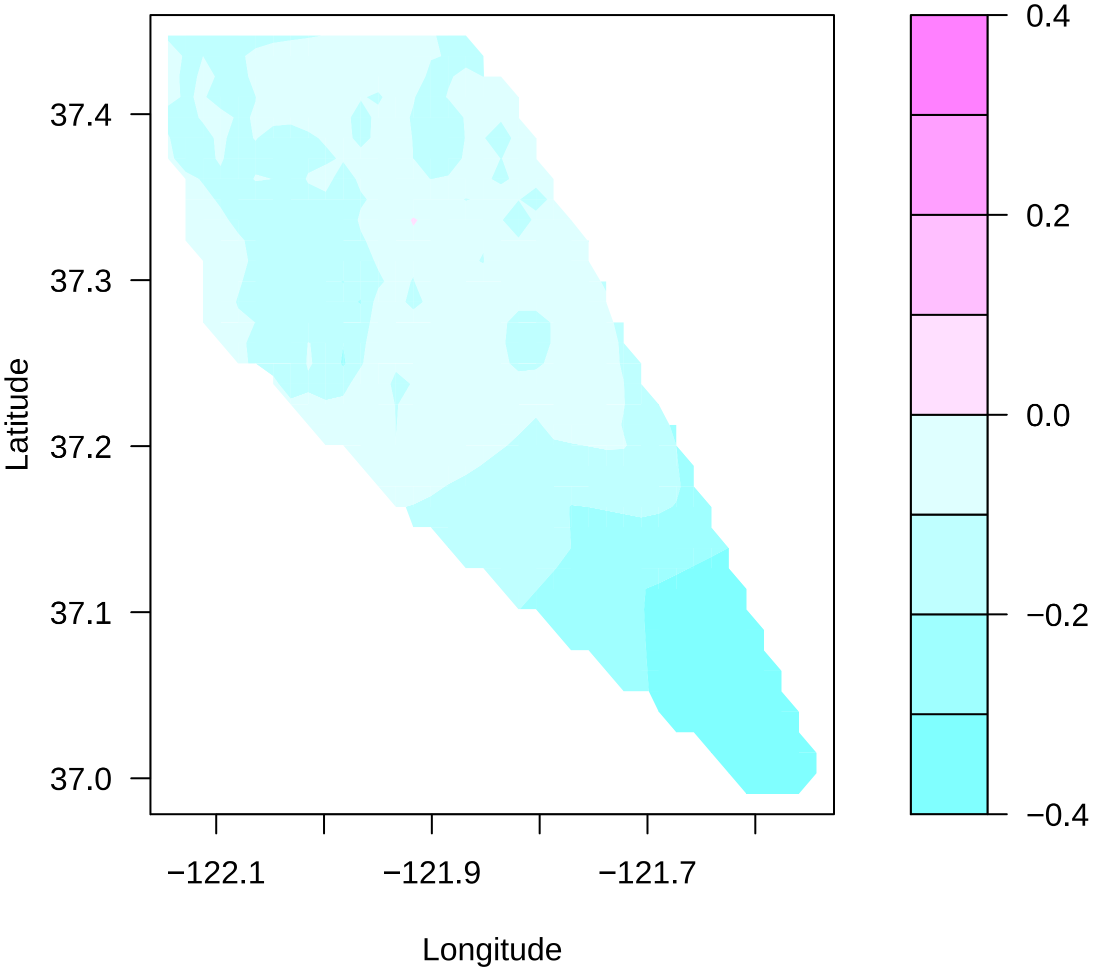


Figure 4.20: Difference in AUC between ULS and SaTScan (ULS-SATSCAN) with RR=3 for simulated circular clusters. It shows that ULS is worse than SaTScan almost everywhere except a very tiny region.

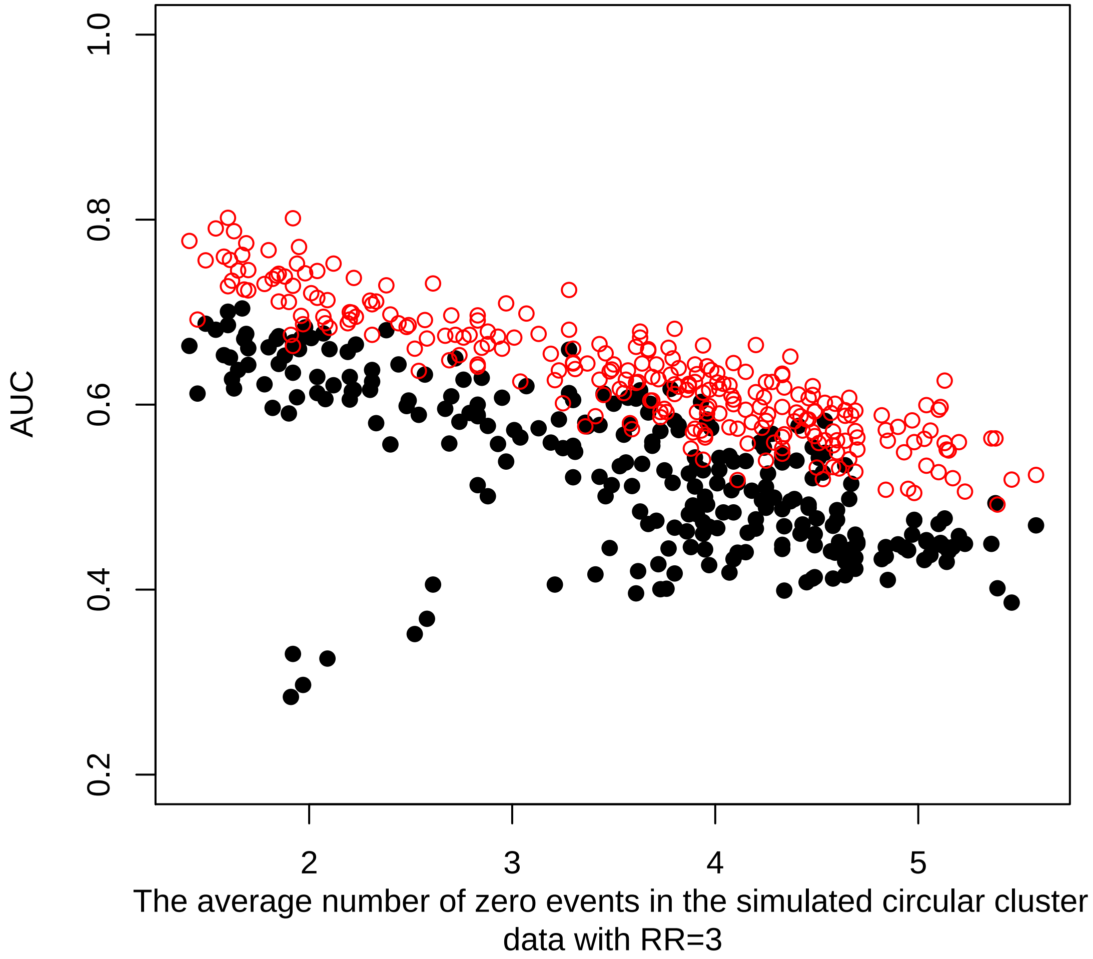


Figure 4.21: Scatter plot of AUC vs. the average number of zero event in the predefined circular cluster for RR=3. Red denotes AUC for SaTScan and black denotes AUC for ULS.

When the relative risk of simulated data is increased to five, AUC values are improved in general for both the SaTScan and the ULS as shown in Figure 4.22 and Figure 4.23. The SaTScan approach performs well across the region. It reveals the similar increasing AUC trend with the rising people at risk for both the SaTScan and the ULS as when RR=3. The same seven tracts fall outside outstand the trend again.

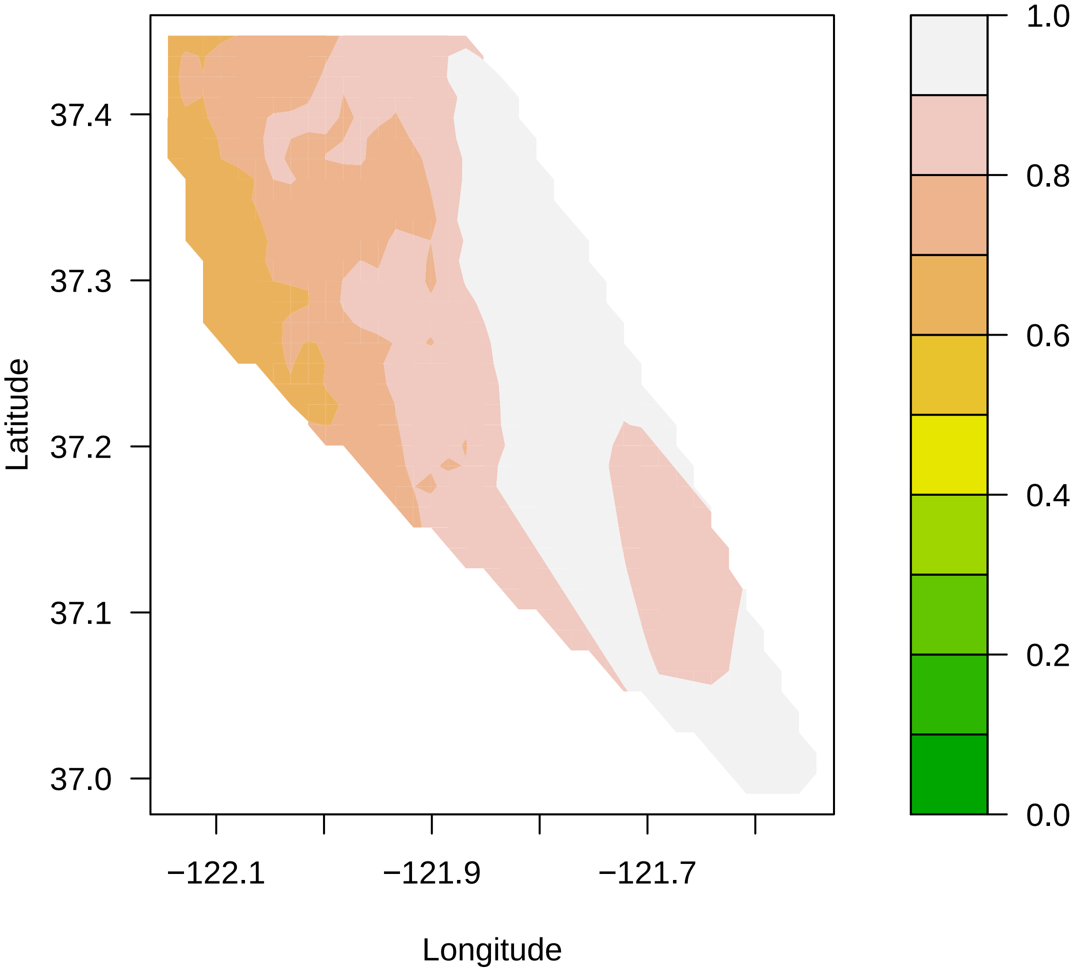


Figure 4.22: Contour map of AUC for SaTScan with RR=5 for simulated circular clusters.

Figure 4.24 shows the difference of AUC between the ULS and the SaTScan based on simulated circular clusters with RR=5. In every location, the SaTScan approach out performed the ULS. We further examined the empirical mean number of zero events in the predefined seven tract cluster. Figure 4.25 shows the number of zero events affects the preformance of the ULS dramatically.

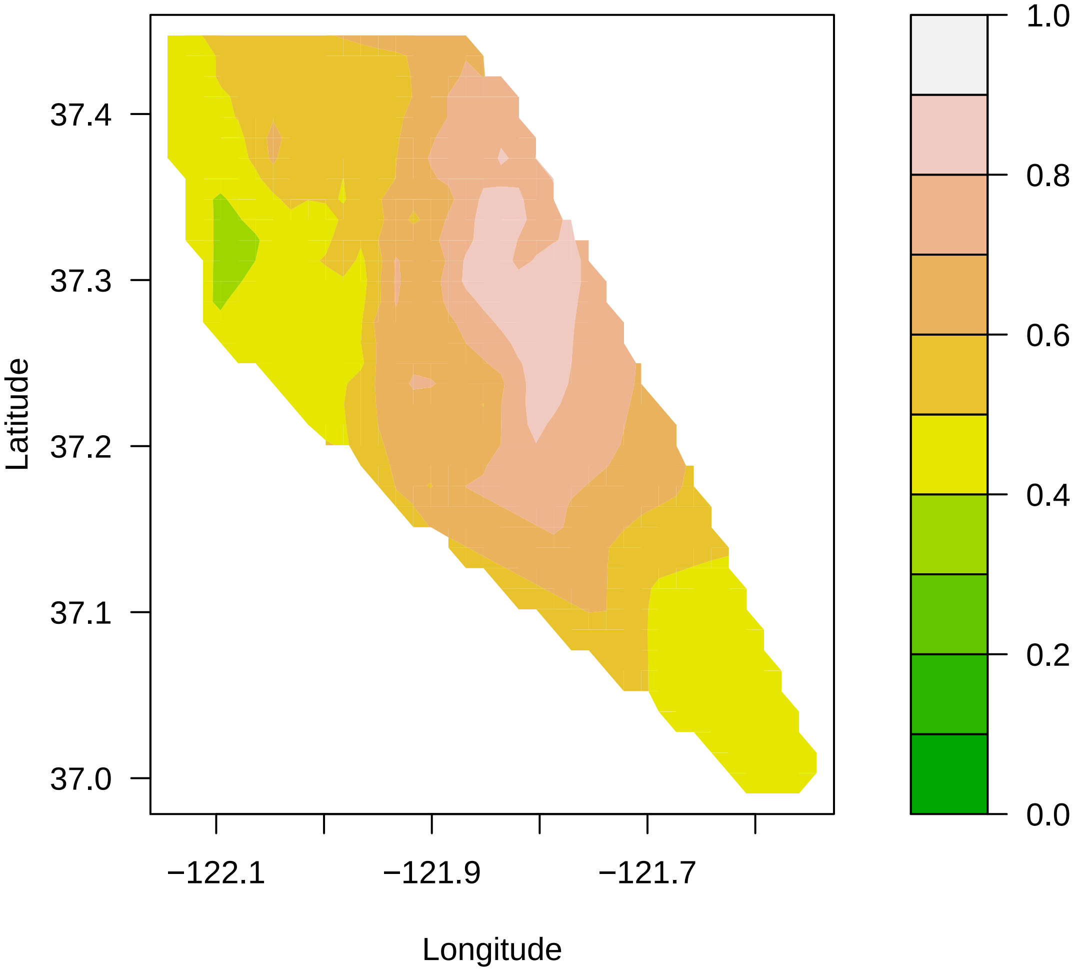


Figure 4.23: Contour map of AUC for ULS with RR=5 for simulated circular clusters.

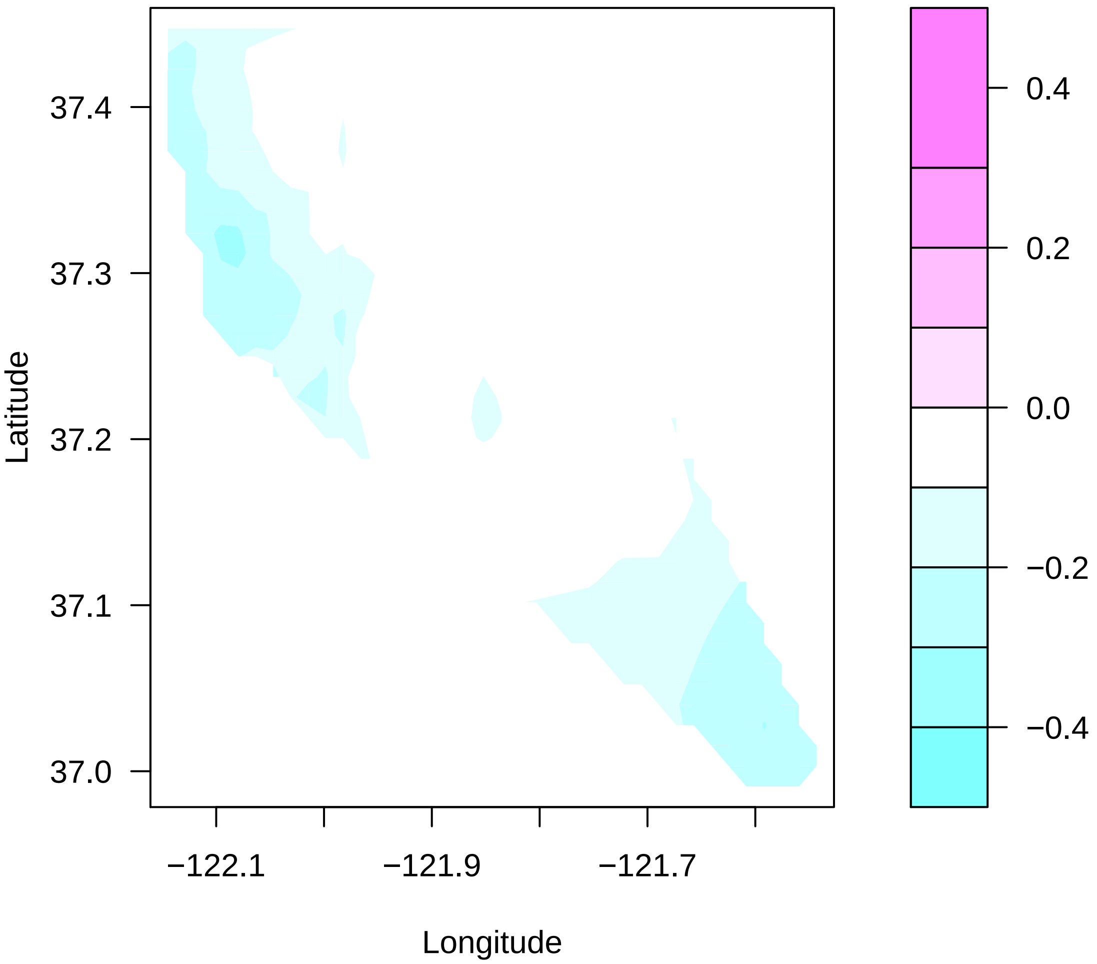


Figure 4.24: Difference of AUC between ULS and SaTScan (ULS-SATSCAN) for RR=5 for simulated circular clusters.

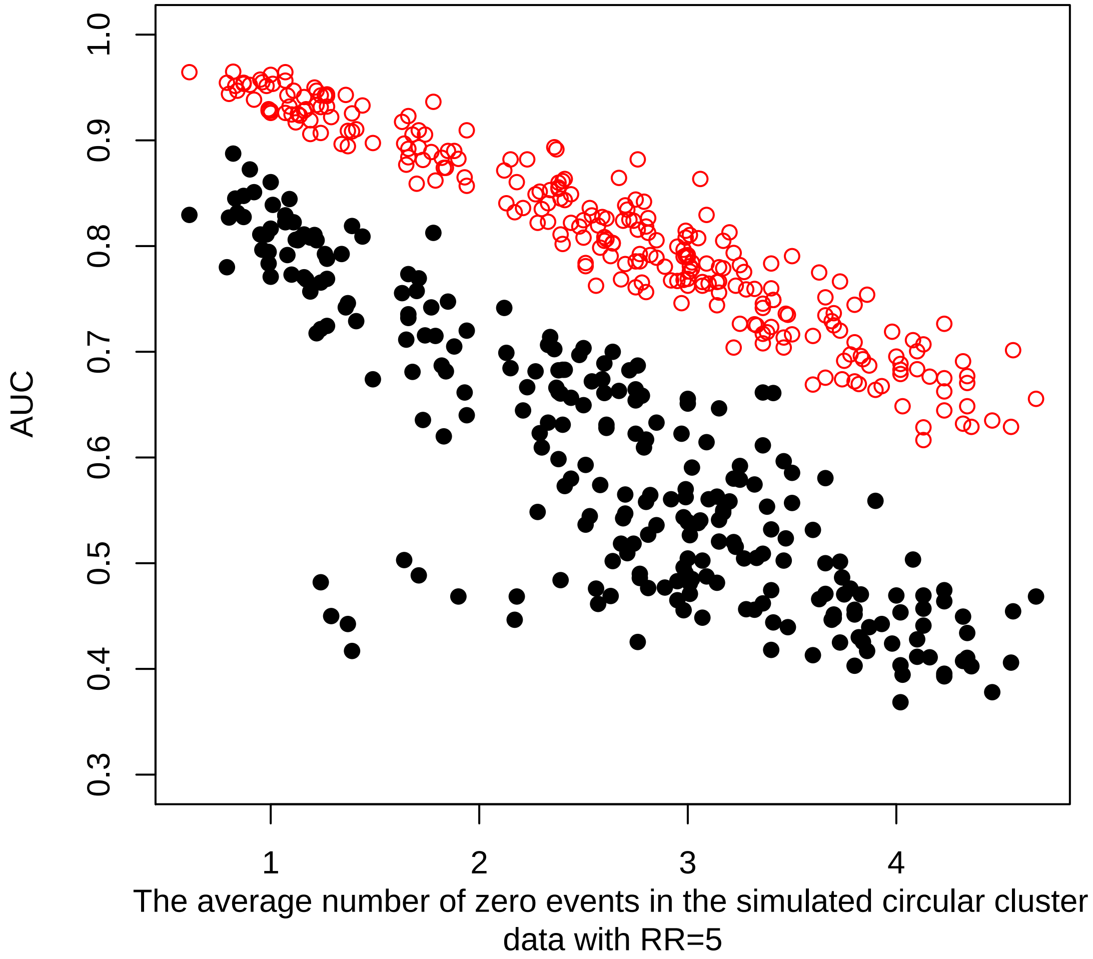


Figure 4.25: Scatter plot of AUC vs. the average number of zero event in the predefined circular cluster for RR=5. Red denotes AUC for SaTScan and black denotes AUC for ULS.

4.5.4 Case II: Closest Neighbors Cluster

Since the Upper Level Set scan method preforms better for irregular shaped cluster, we might expect ULS to outperform SaTScan when simulated some clusters based on their neighbors as describe in the previous section. Figure 4.26 shows the contour plot based on our local AUC values for the connected neighbors cluster with $RR = 3$ by the SaTScan approach. We note that the AUC varies fairly dramatically with location ranging from 0.5 to 0.8, and performs worse than for circular cluster simulations as shown in Figure 4.27 in most locations. Therefore, SaTScan performs better when the underlying cluster is circular than when it follows irregular shapes.

Figure 4.28 shows the contour plot based on our local AUC values for the connected neighbors cluster with $RR = 3$ for the ULS. We note that the AUC varies fairly widely with location ranging from 0.3 to 0.7. It performs better than for circular simulations as shown in Figure 4.29 in some locations, but worse in other locations. It still reveals a general trend of increasing AUC with increasing number of live births except for five tracts at the far south corner since their choice of neighbors are limited in the corners. The only tract in this corner that performs better is the one has its neighbors beyond that corner. When comparing with the SatScan results, in most locations the ULS is performance worse as shown in Figure 4.30, but the regions that outperform the SaTScan are larger than in the circular situation.

Figure 4.31 shows the number of zero event tracts in the predefined close neighbor cluster affects AUC for both the SaTScan and the ULS, however, it affects the ULS more dramatically especially when the number of zero events are higher than three.

Figure 4.32, Figure 4.33, and Figure 4.34 show the number of live births, the number of cases, and the incidence rate, which is defined as the ratio of the number of cases and the number of live births in the predefined close neighbor cluster, affects both the performance of the SaTScan and the ULS. Furthermore, the ULS has more variability, specially at the lower number of live births in the predefined close neighbor cluster. The ULS performance is much closer to that of SaTScan at higher number of live births for regions in the close neighbor cluster compared to the circular cluster.

In summary, the performance of the ULS is improving with non circle clusters. However, SaTscan still performs better in most locations, and SaTScan follows the population contours more closely.

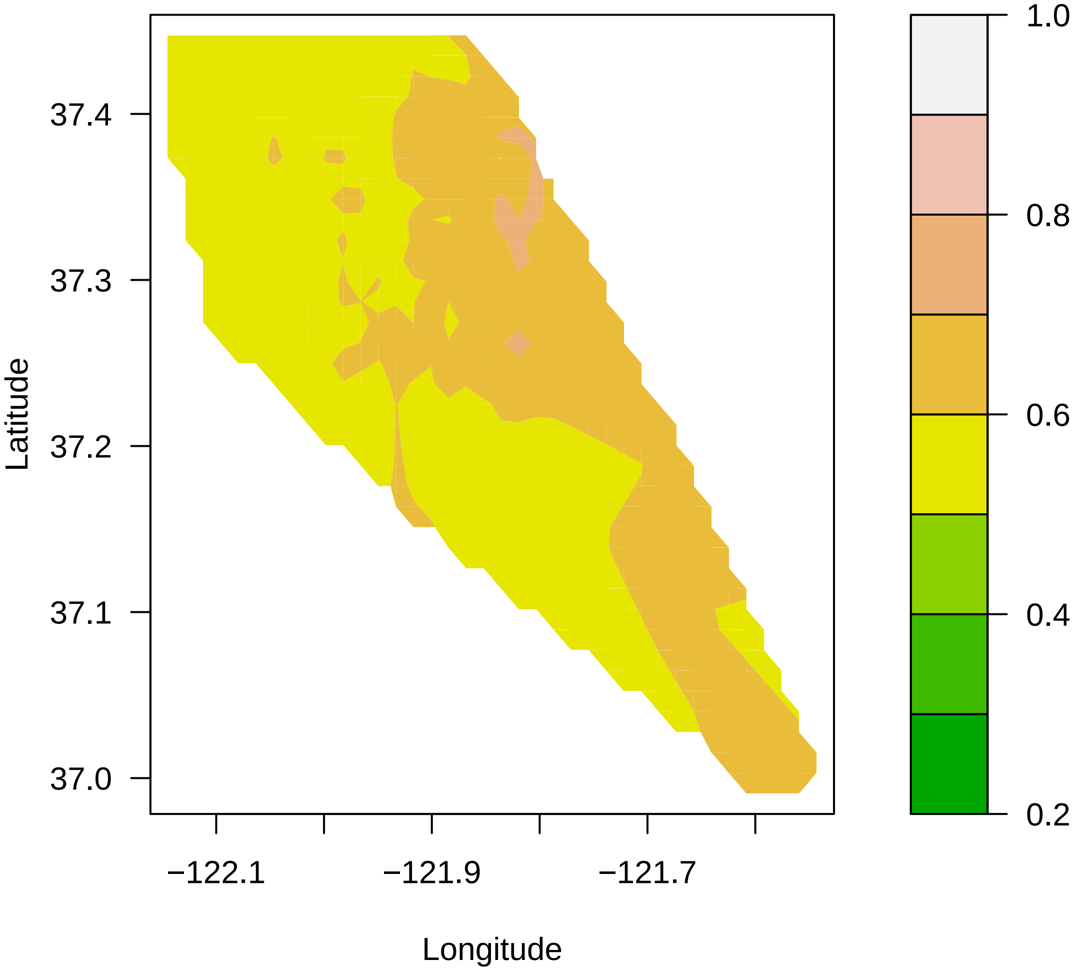


Figure 4.26: Contour map of AUC for SaTScan with RR=3 for simulated close neighbor clusters.

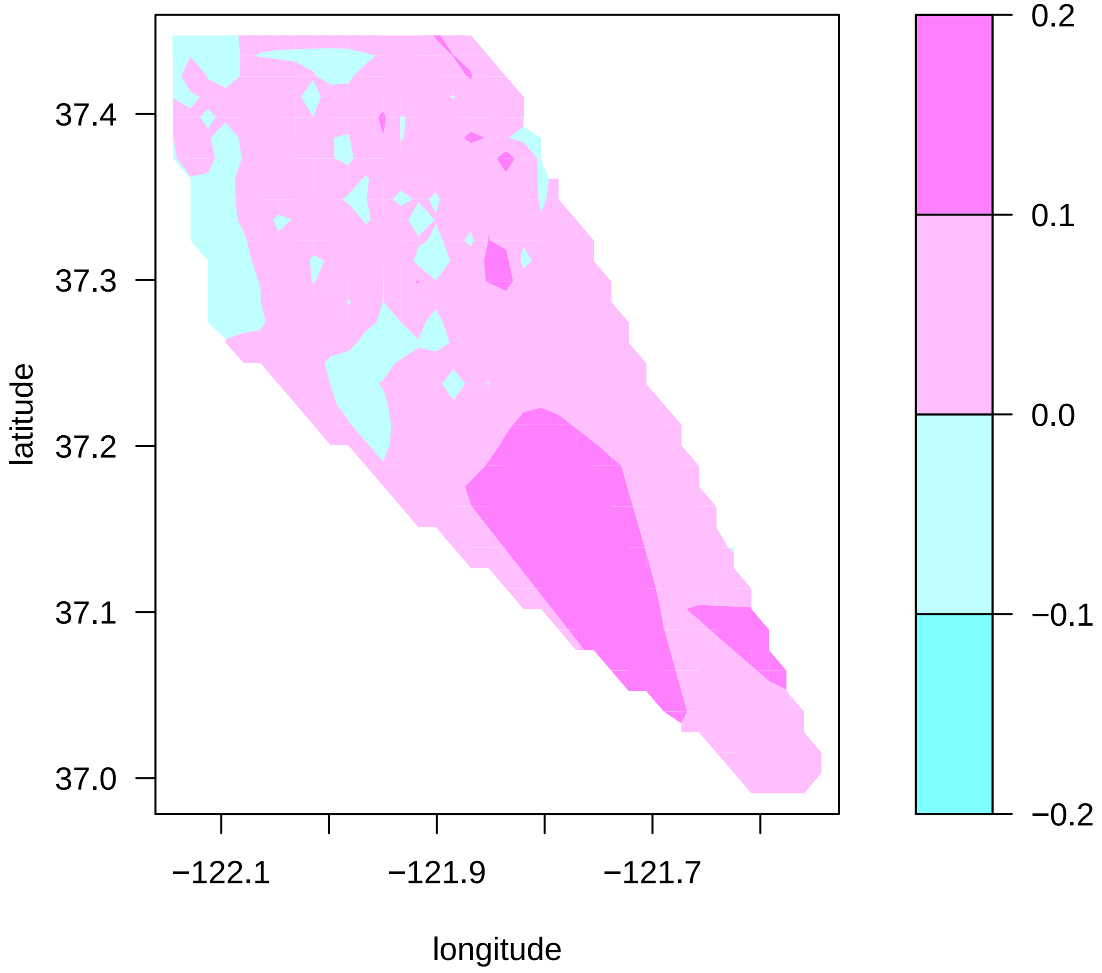


Figure 4.27: Contour map of AUC for difference between circular and non-circular clusters (DIFF = circular-close neighbor) for SaTScan with RR=3.

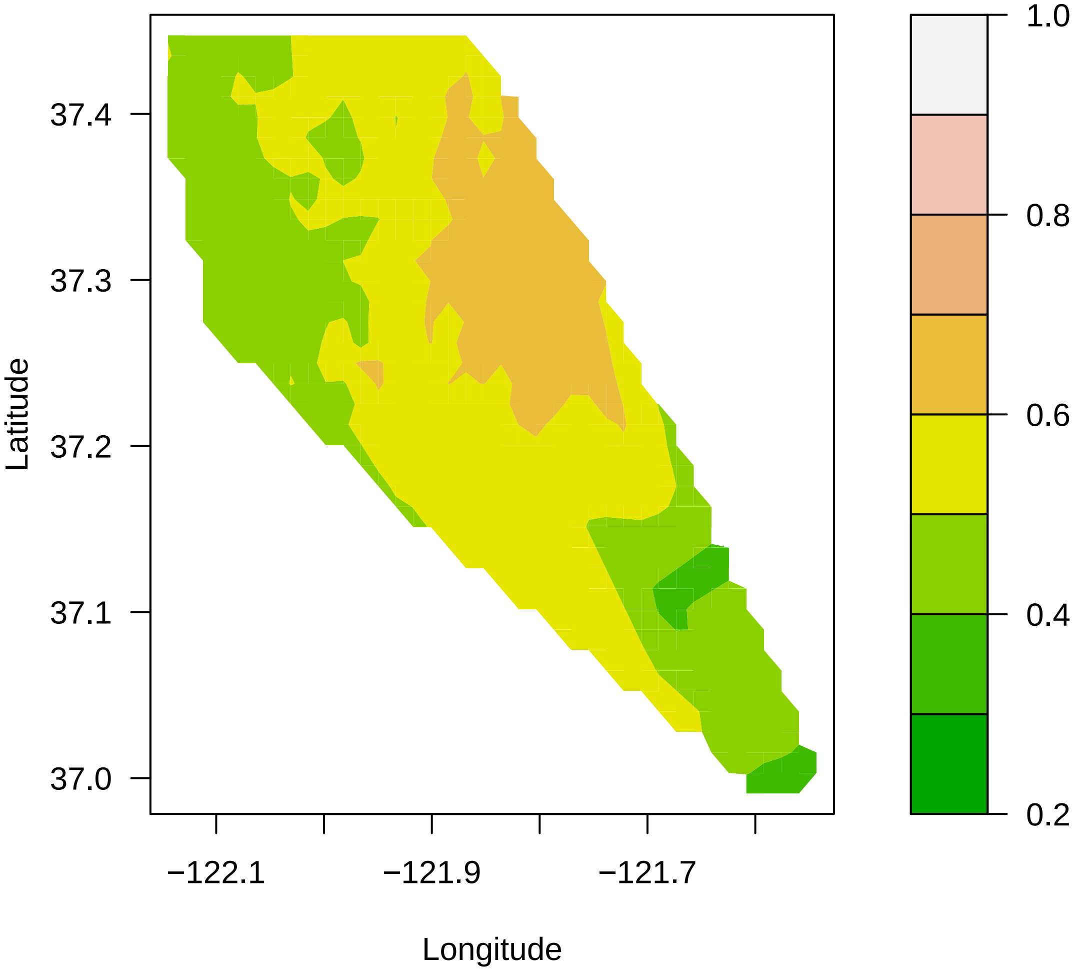


Figure 4.28: Contour map of AUC for ULS with RR=3 for simulated non circular clusters.

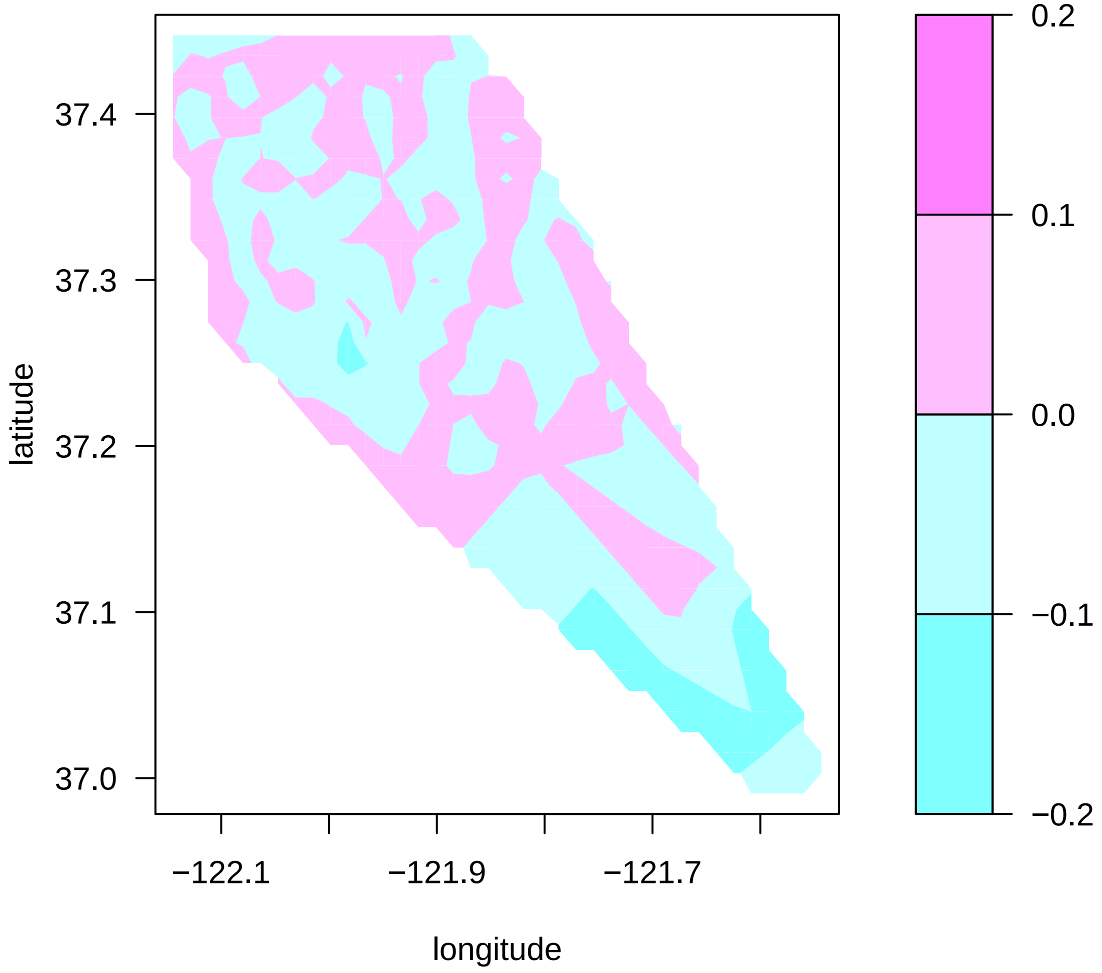


Figure 4.29: Difference of AUC for ULS between circular and non-circular clusters (circle - non-circular) with RR=3.

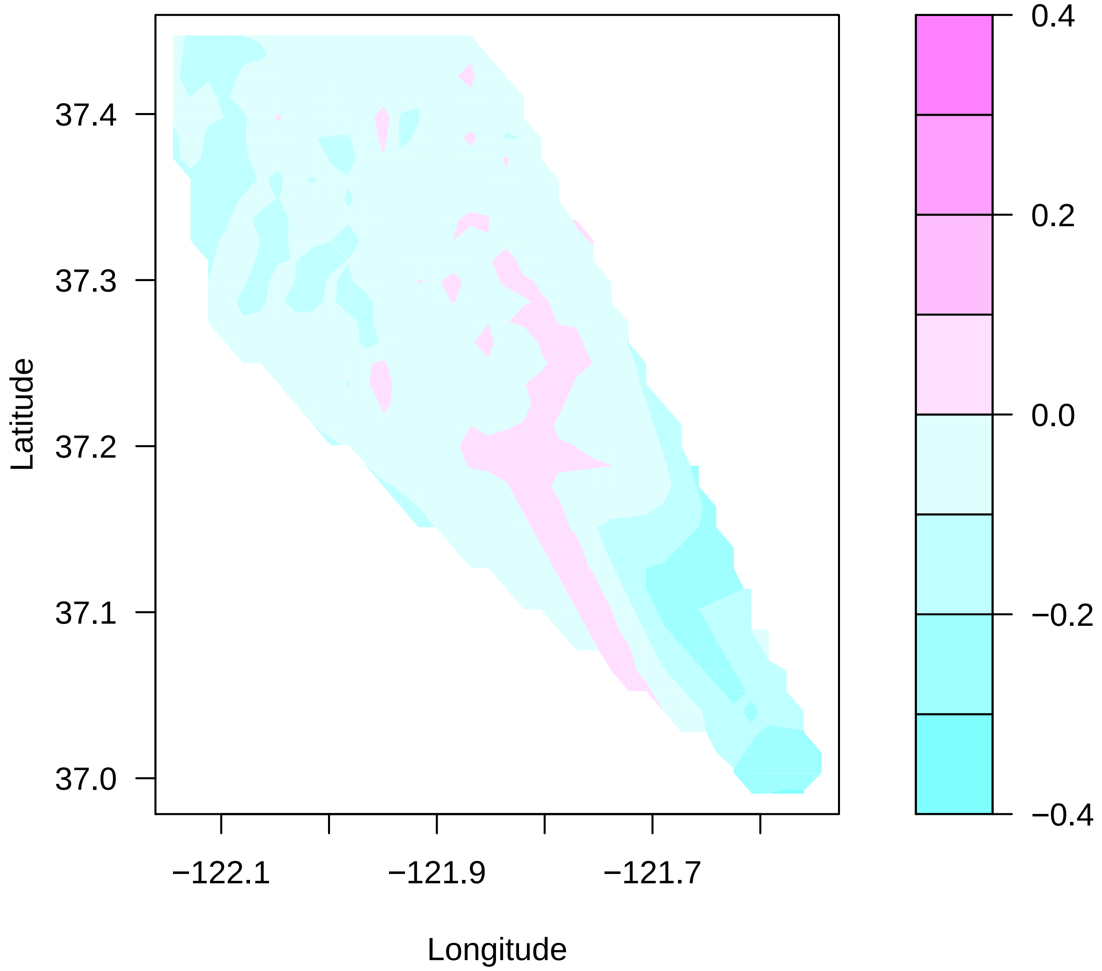


Figure 4.30: Difference of AUC between ULS and SaTScan (ULS-SATSCAN) with RR=3 for simulated non circle clusters.

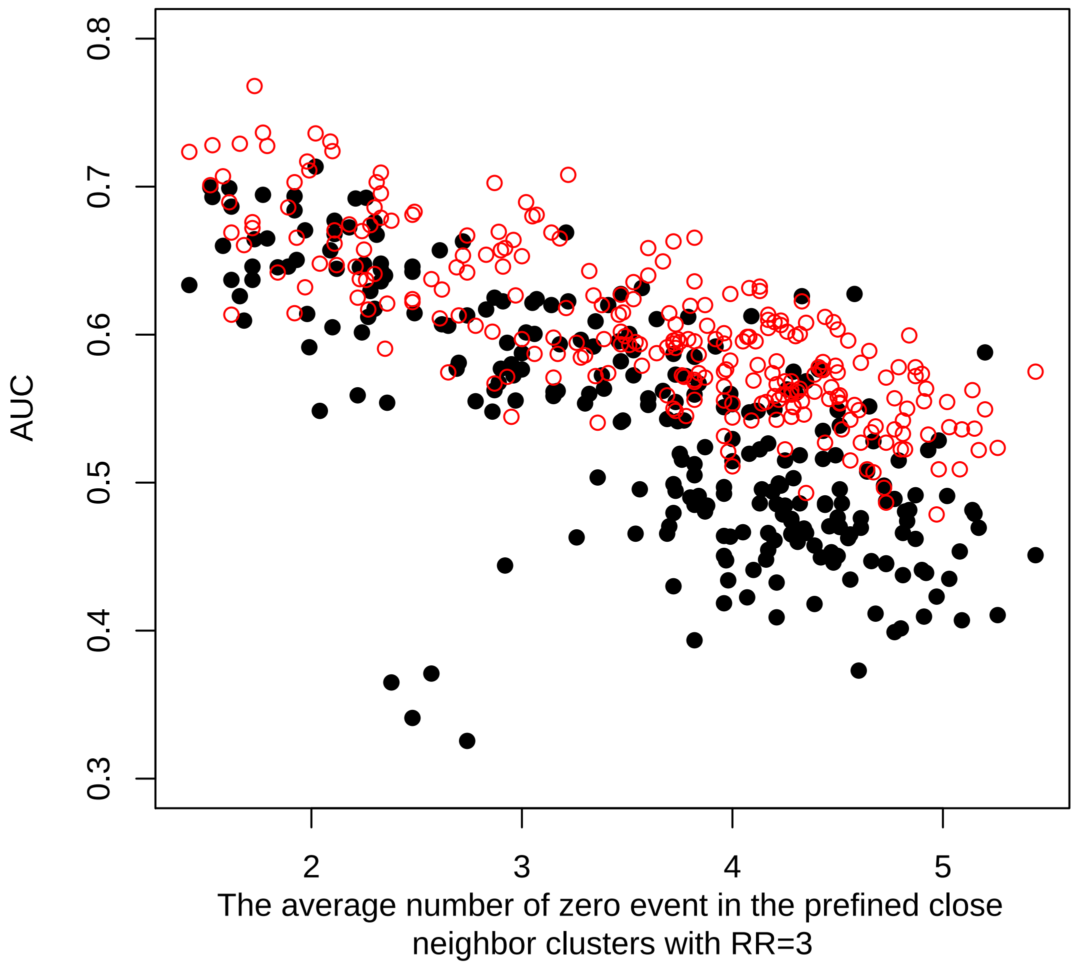


Figure 4.31: Scatter plot of AUC vs. the average number of zero event in the predefined non-circular cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

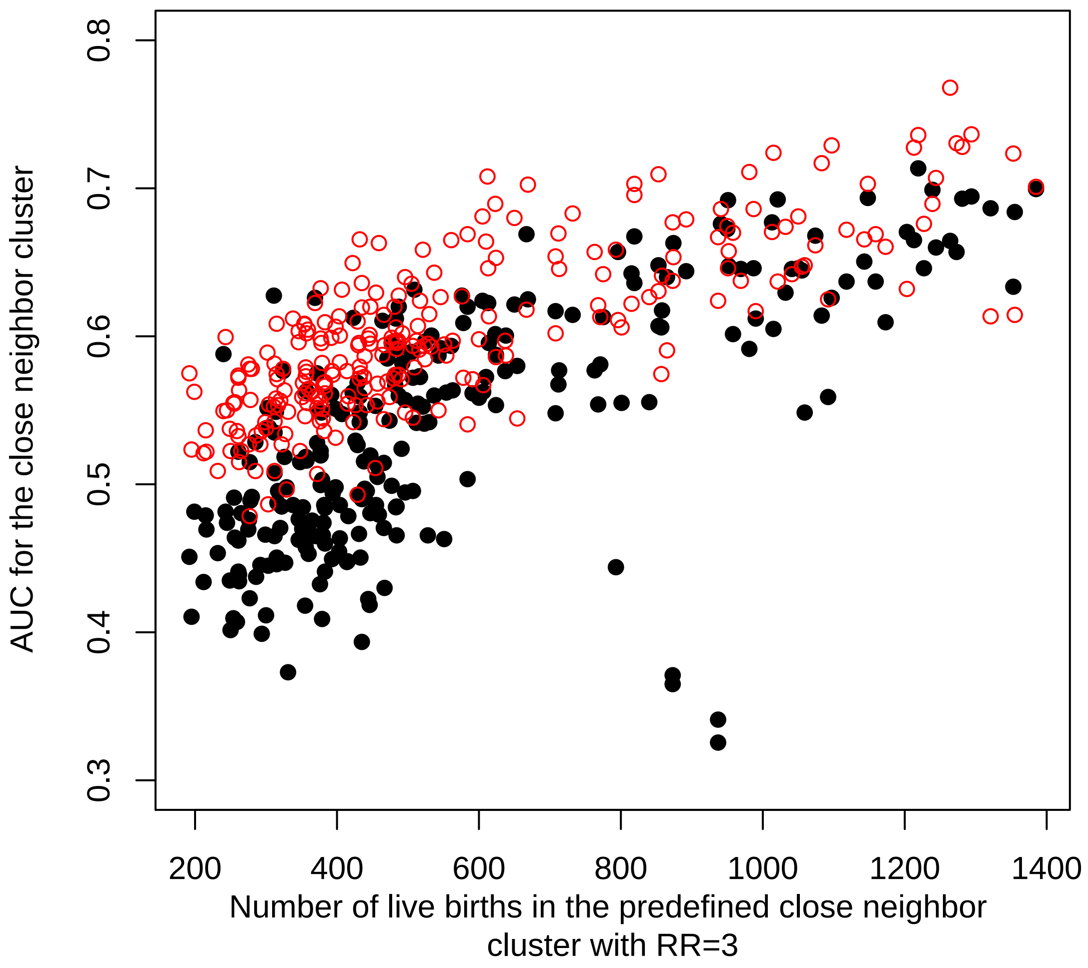


Figure 4.32: Scatter plot of AUC vs. the number of live births in the predefined close neighbor cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

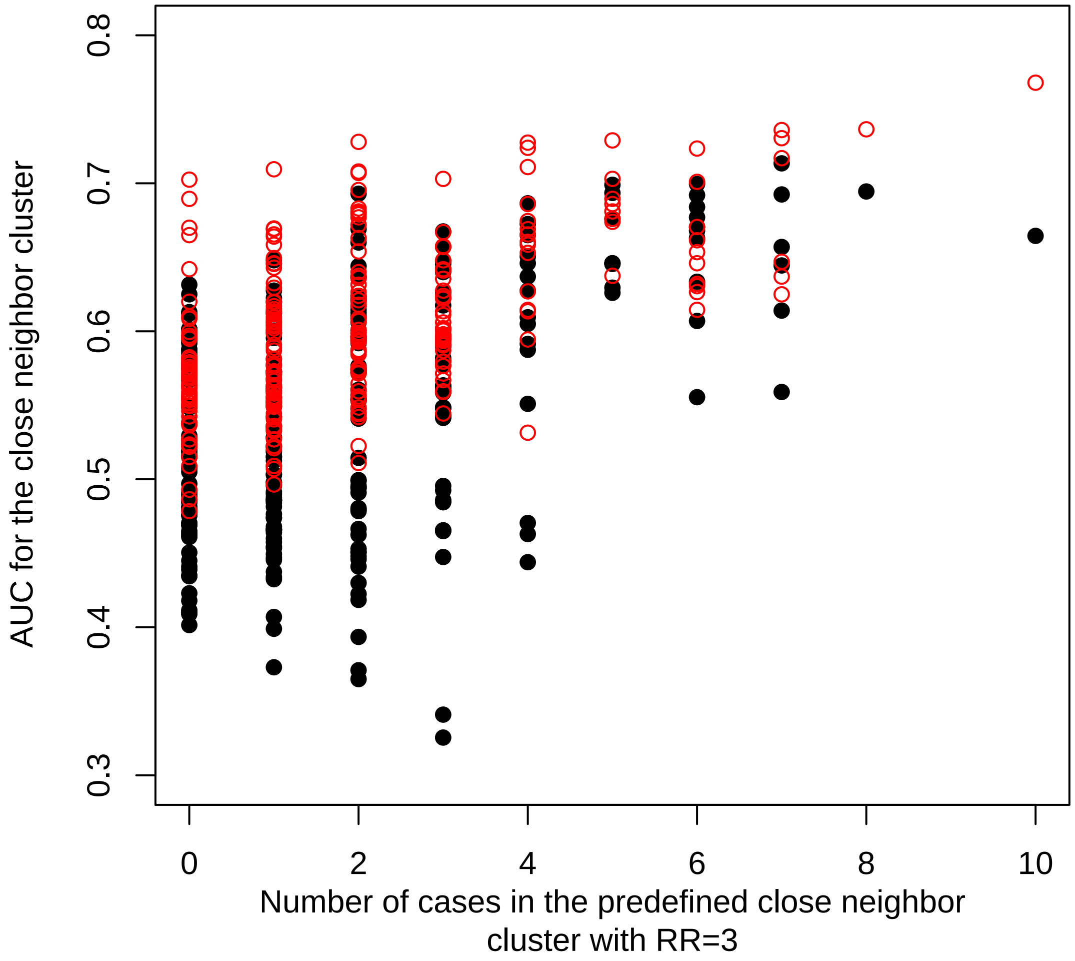


Figure 4.33: Scatter plot of AUC vs. the number of cases in the predefined close neighbor cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

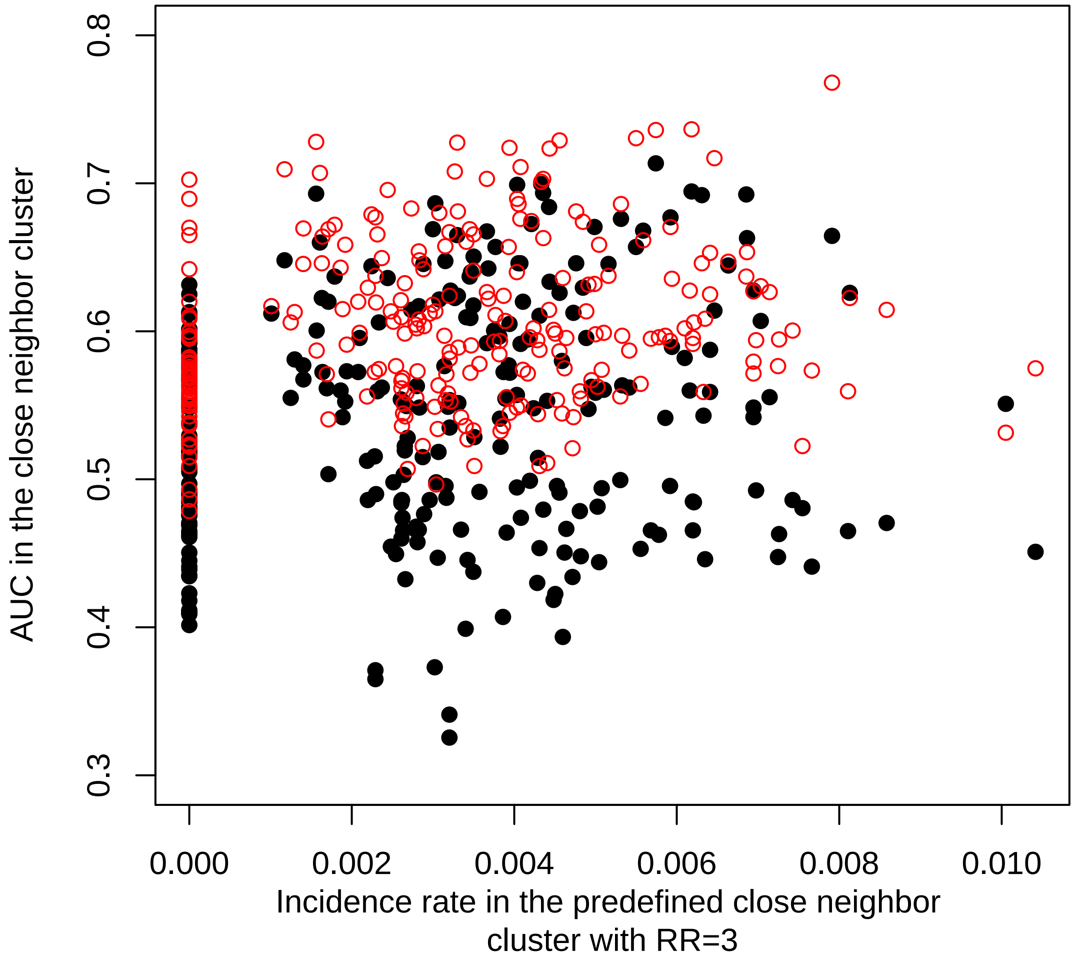


Figure 4.34: Scatter plot of AUC vs. the incidence rate in the predefined close neighbor cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

4.5.5 Case III: Elongated Neighbors Cluster

Since neighbors may be still close to each other, we next purposely selected the furthest neighbor to make the cluster elongated. We applied ULS and SaTscan to this type of cluster to test if it favors upper level set scan method. Figure 4.35 shows the contour plot based on our local AUC values for the farther elongated clusters at $RR = 3$ by the SaTScan. We note that the AUC varies from 0.4 to 0.7. It performs worse than the underlying cluster is circular in most locations as shown in Figure 4.36.

We applied the ULS to the same simulated data used for SaTScan, and Figure 4.37 shows the contour plot. We note that AUC varies dramatically with location, and the area that it performs better than SaTScan in enlarged over that in the previous two cases as shown in Figure 4.38. In addition, in many locations, especially in the middle section it performs better than when the underlying cluster is circular as shown in Figure 4.39. This confirms that the ULS is better at detecting irregular shaped clusters.

Figure 4.40 shows that the number of zero events affects the performance of the ULS dramatically. Figure 4.41, Figure 4.42, and Figure 4.43 show the number of live births, the number of cases, and the incidence rate, which is defined as the ratio of the number of cases and the number of live births in the predefined elongated cluster, affect both the performance of the SaTScan and the ULS. Furthermore, the ULS has more variability than SaTScan.

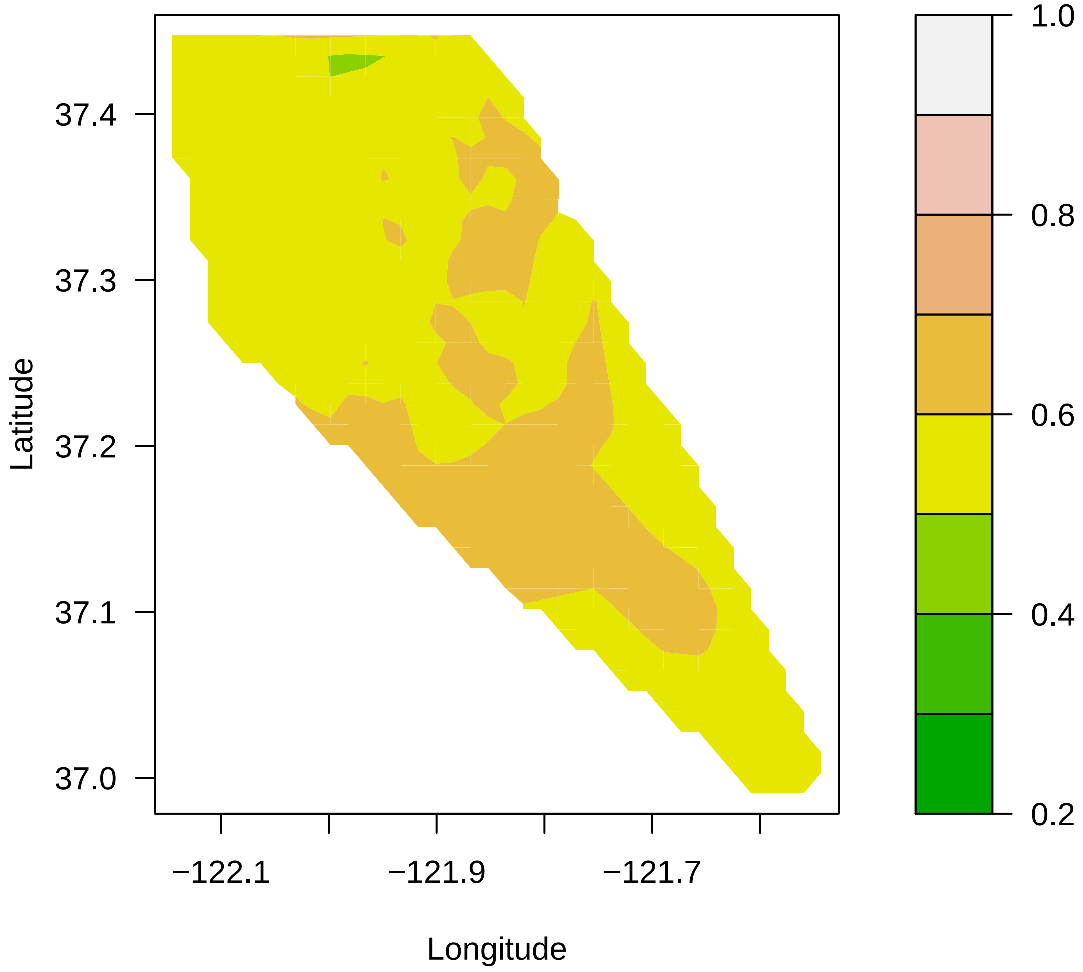


Figure 4.35: Contour map of AUC for SaTScan with RR=3 for simulated non circle elongated clusters

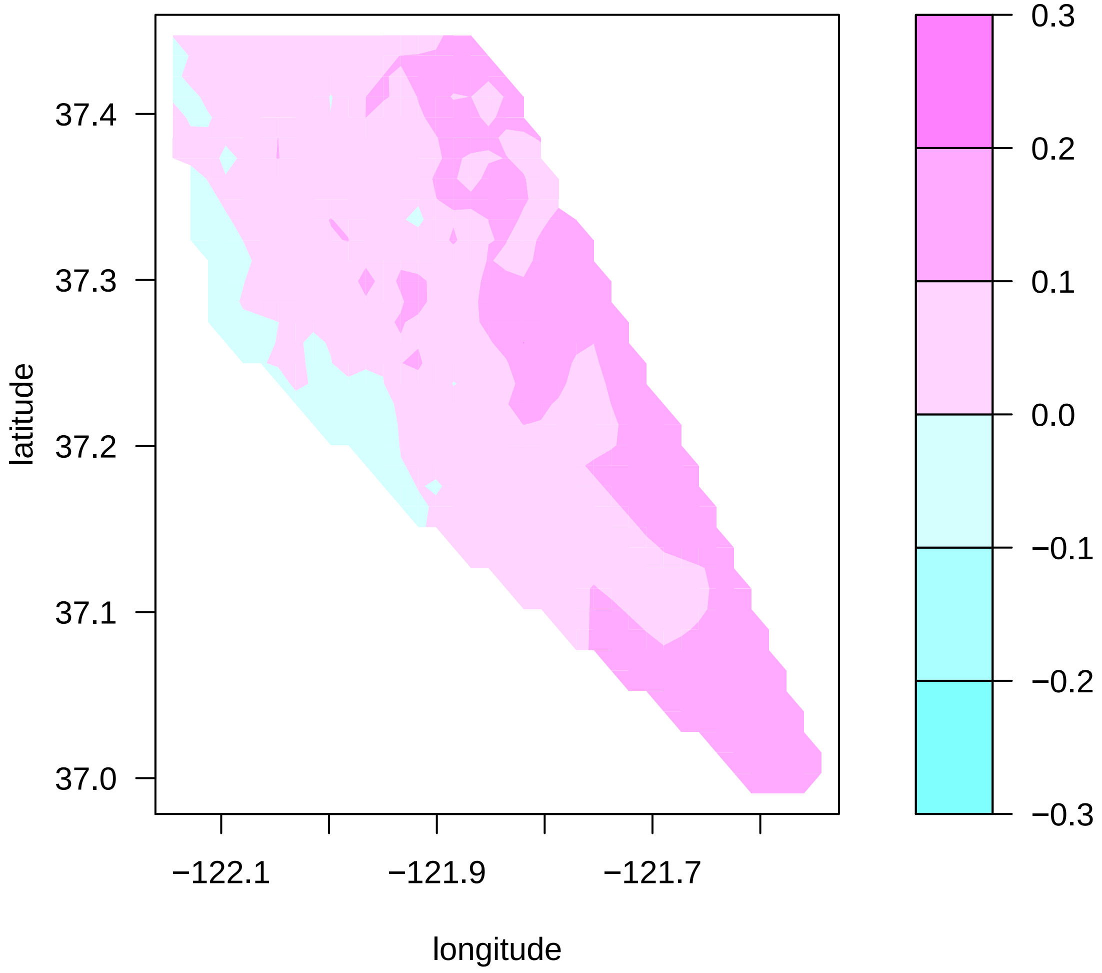


Figure 4.36: Difference of AUC for SATSCAN between circular and elongated clusters (circular - elongated) with RR=3.

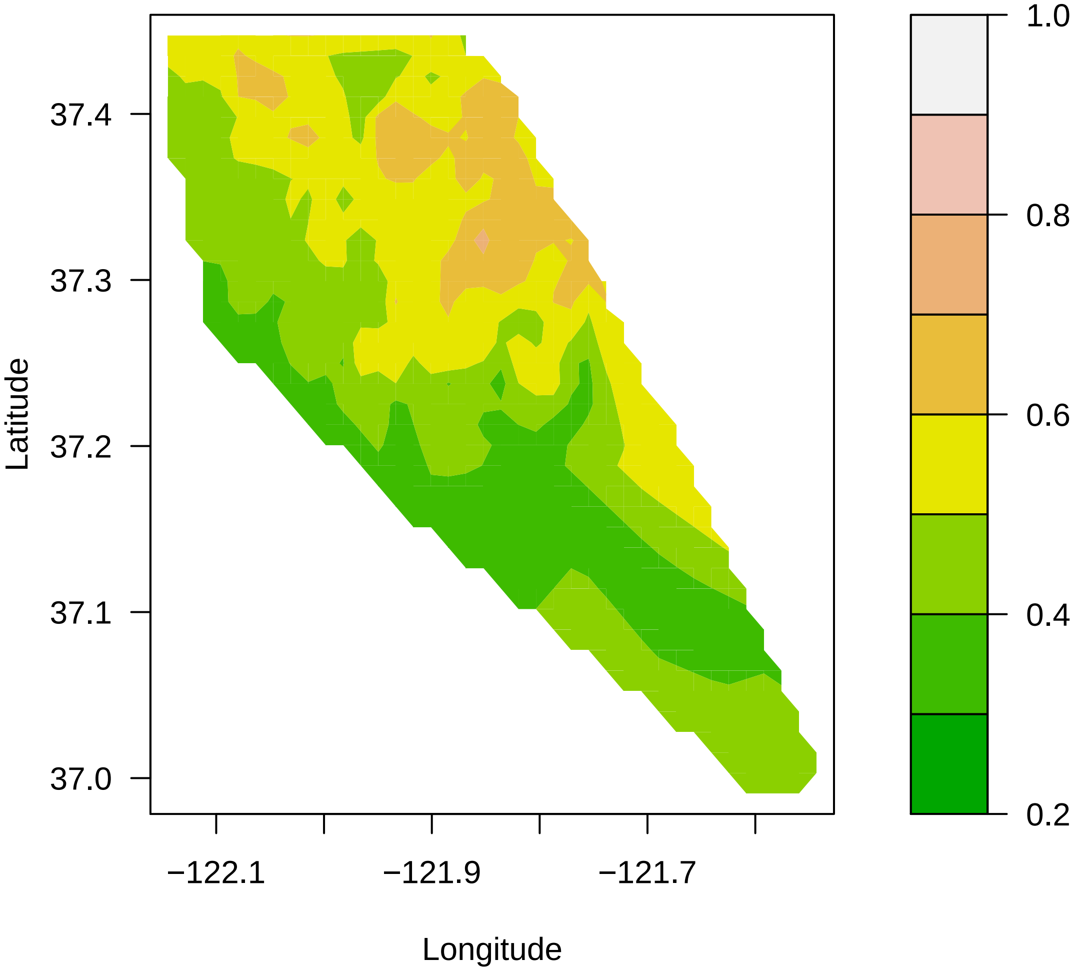


Figure 4.37: Contour map of AUC for ULS with RR=3 for simulated elongated clusters

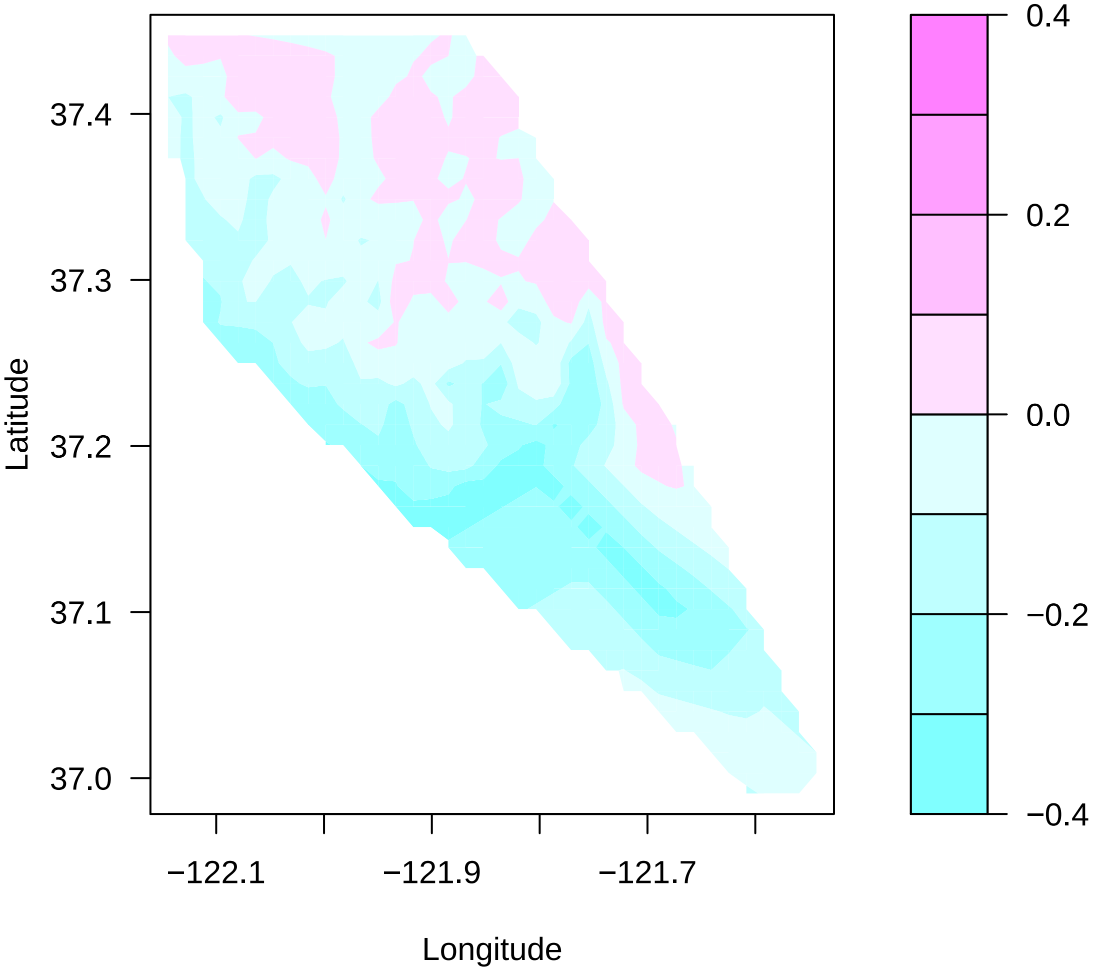


Figure 4.38: Difference of AUC between ULS and SaTScan (ULS-SATSCAN) with RR=3 for simulated elongated clusters.

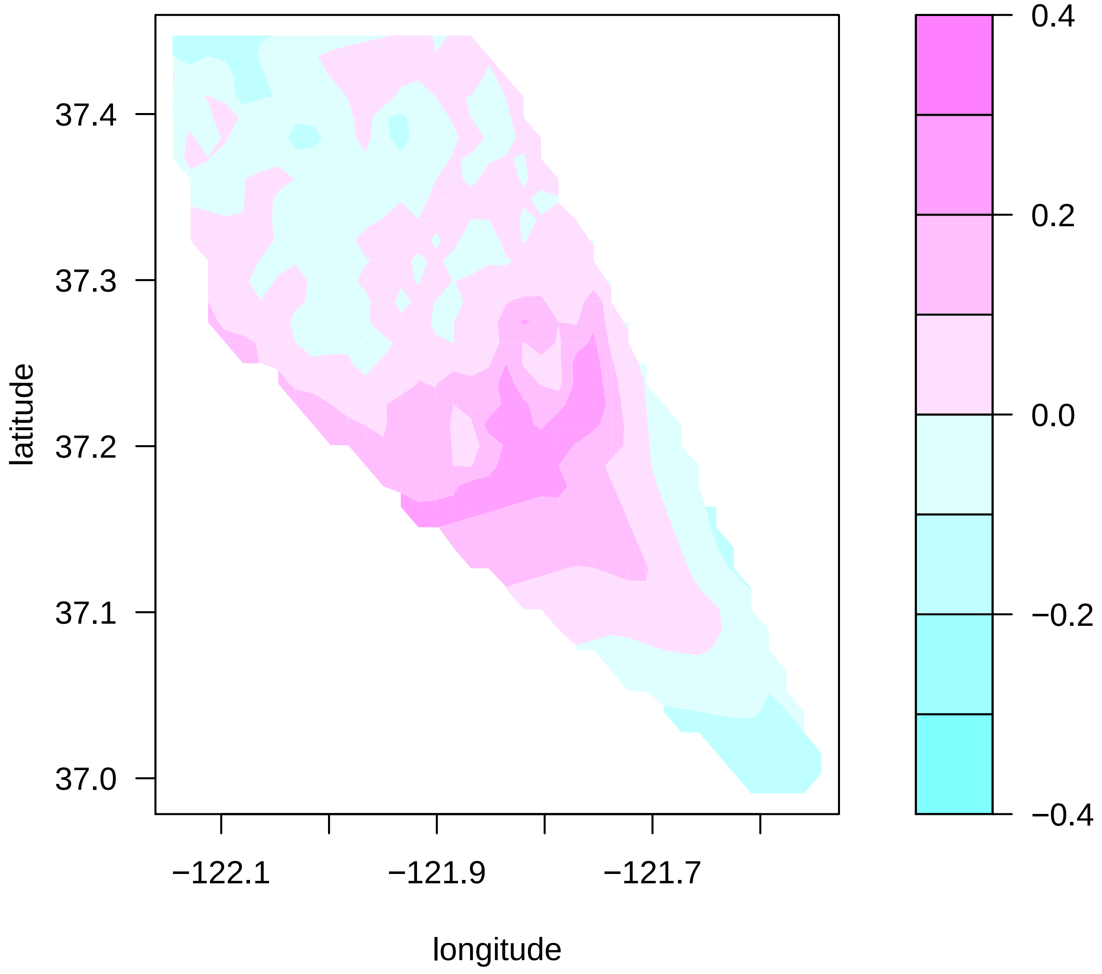


Figure 4.39: Difference of AUC for ULS between circular and elongated (circular - elongated) with RR=3 for simulated clusters.

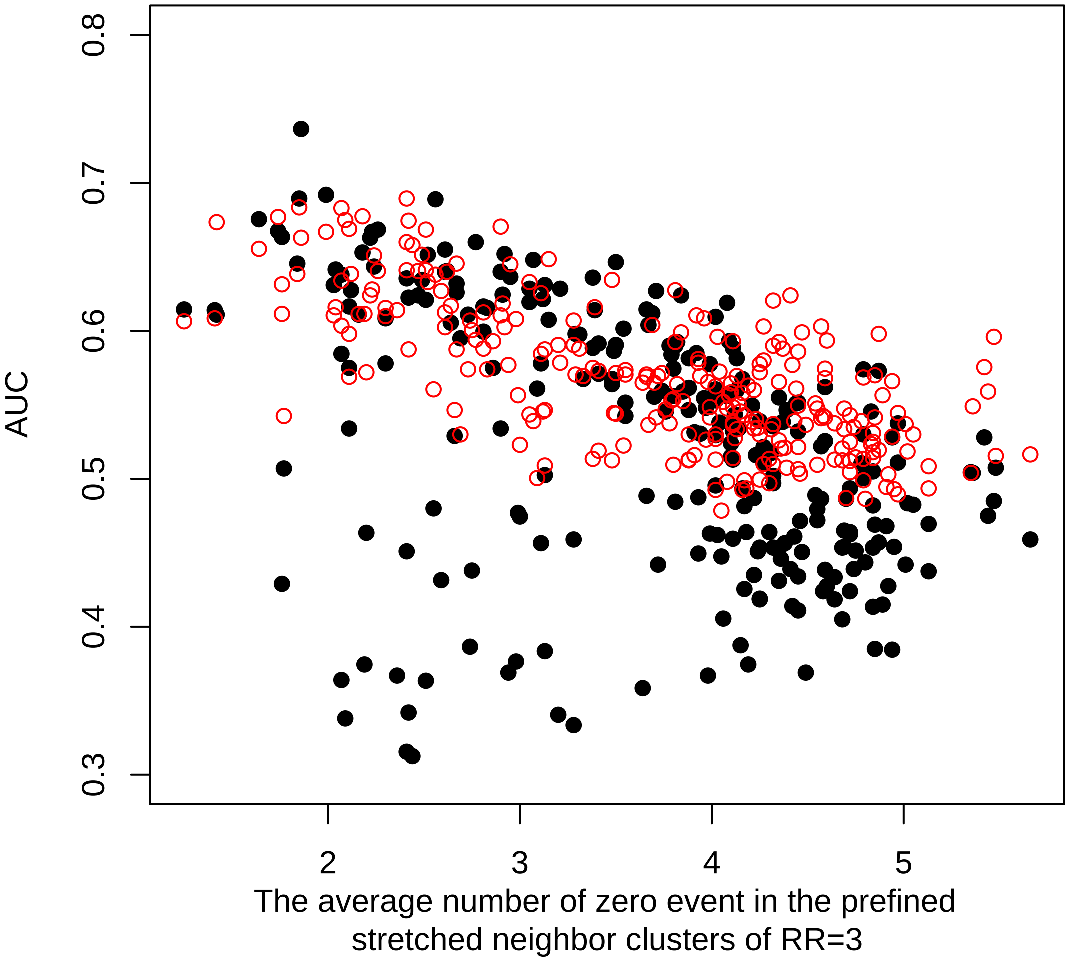


Figure 4.40: Scatter plot of AUC vs. the average number of zero event in the predefined elongated cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

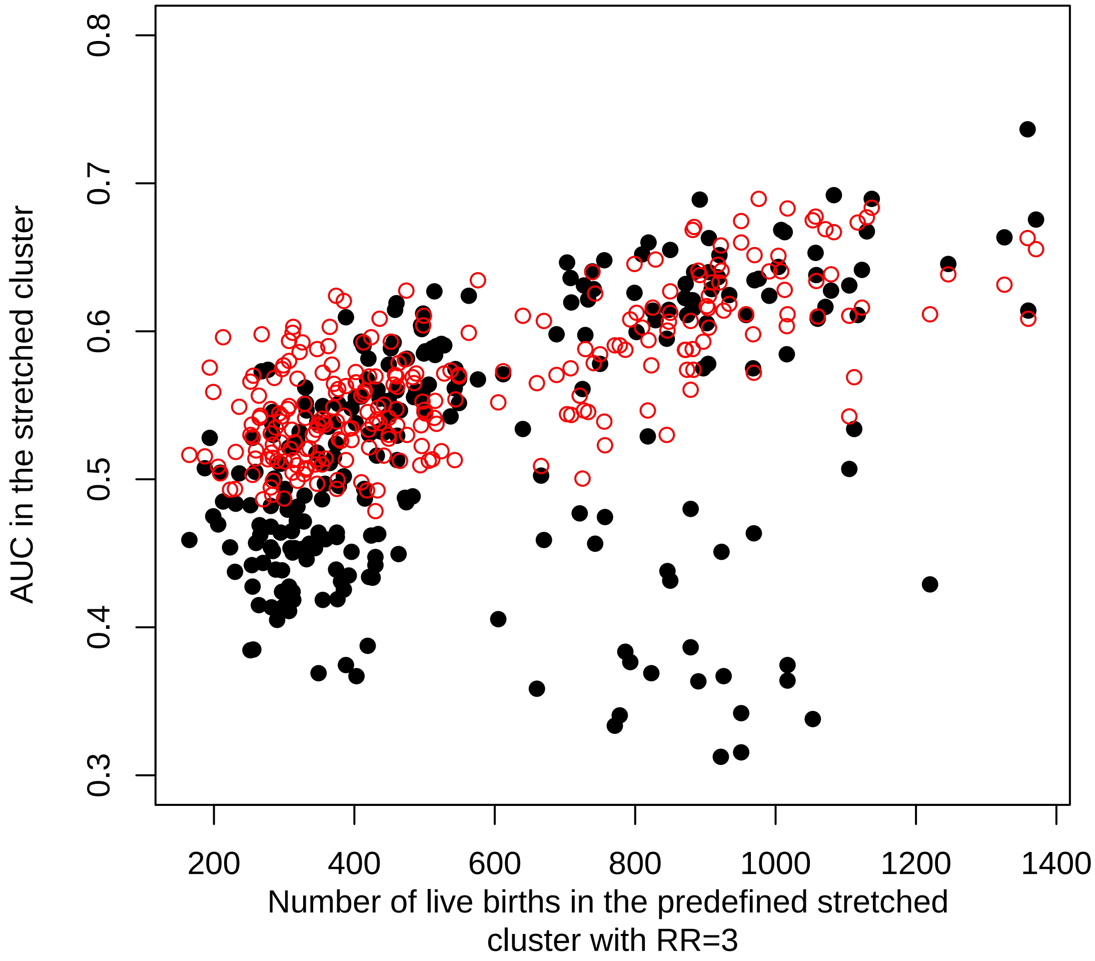


Figure 4.41: Scatter plot of AUC vs. the number of live births in the predefined elongated cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

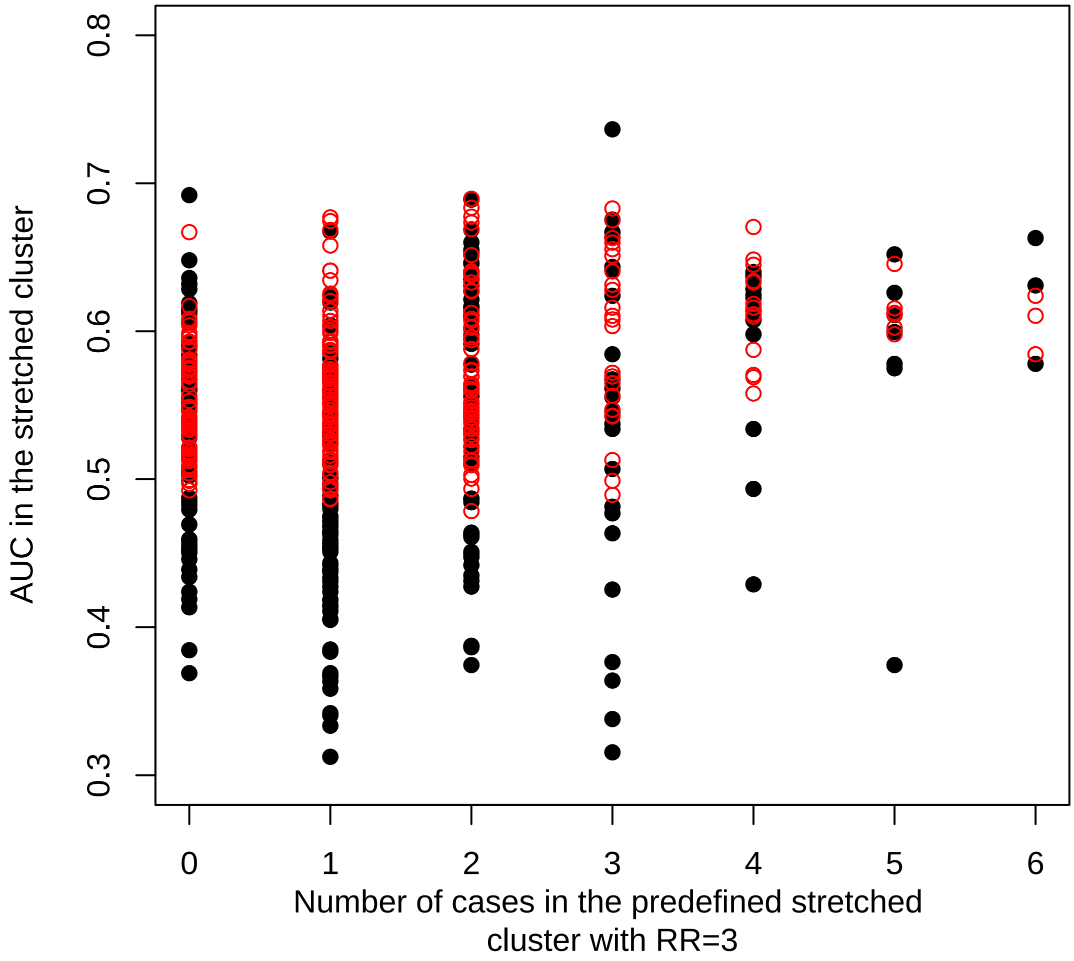


Figure 4.42: Scatter plot of AUC vs. the number of cases in the predefined elongated cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

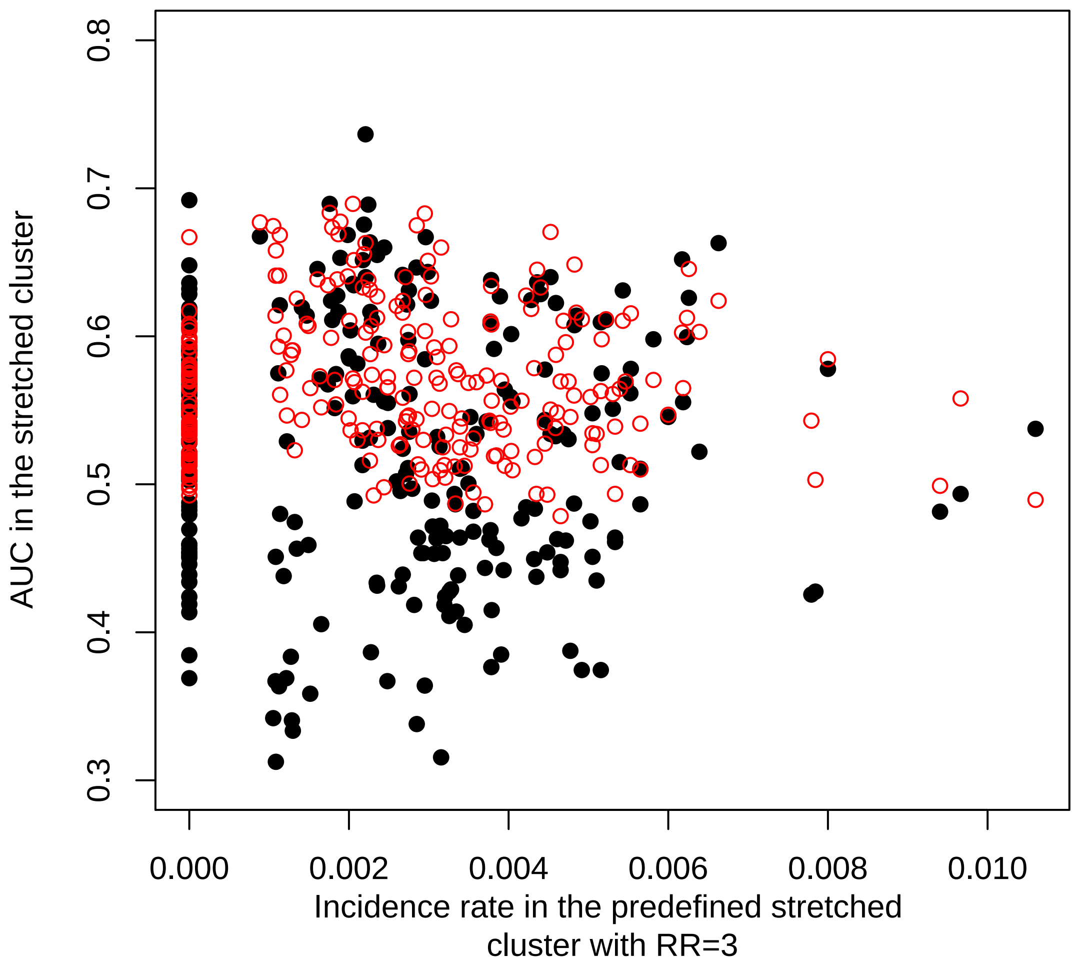


Figure 4.43: Scatter plot of AUC vs. the incidence rate in the predefined elongated cluster with RR=3. Red denotes SaTScan, and black denotes ULS.

4.6 Conclusions

In this chapter, we applied both the SaTScan and the Upper Level Set scan to the Santa Clara data set, and computed area under the receiver operating characteristic curve for these types of clusters by Monte Carlo simulations. The results reveal the SaTScan performs better if the cluster is compact, and the Upper Level Set approaches offer improved performance when clustering is irregularly shaped. Therefore, when irregular cluster is suspected, the ULS method should be considered. In addition, the performance of cluster detection methods can be highly variable in space, often ranging from completely ineffective (AUC < 0.5) to extremely accurate.

Chapter 5

Spatial Measures of Performance: Regional Point Data for Cases and Controls

In this chapter, we investigate the problem of cluster detection in regional point data. The spatial performance of Kernel Density method and Generalized Additive Models are compared.

5.1 Introduction

Quantification of spatial pattern is of interest in a variety of disciplines. For example, detection of disease outbreaks often involves spatial analysis. In archaeology, identification of significant local differences in artifact or settlement patterns can provide a clue to underlying behavioral drivers operating in the past. As in previous chapter, our goal is to derive methods to evaluate the statistical performance of spatial surveillance systems. Spatial surveillance often involves a search for clusters (2; 12). A cluster refers to areas of abnormal observations, such as those with a higher incidence rate, compared to other areas in a particular time period, in space or both (12; 61).

Methods for detecting disease clusters provide tools for exploring data, but as noted earlier, causal effects are typically drawn by further research (28).

In this chapter, we develop an approach to assess and compare performance of two cluster detection methods and illustrate the approach on archaeological data involving Anasazi settlement sites in northeastern Arizona. The Anasazi (or Ancestral Pueblo peoples) were an ancient population of Native Americans who lived in the southwestern United States including Arizona, Colorado, New Mexico and Utah (13; 36; 69). There are thousands of Anasazi archaeological sites, and we focus on the sites studied by Peabody Coal Company's Black Mesa archaeological project, one of the largest archaeological projects in the American Southwest, and conducted in northeastern Arizona from 1967 to 1983. In the following sections, we investigate whether the pattern of late sites dated between 950 and 1050 A.D., which spans the time from rapid population growth to the abandonment of settlements, are different from those early sites dated between 850 and 949 A.D. occuring prior to the sudden population growth. Our data include 100 early sites and 389 late sites. As the majority of the Anasazi in Black Mesa lived in small groups as farmers (13), we use geographic information systems to compute distance measures between sites and local streams and stream beds around those sites to explore the effects of proximity to seasonal water sources on patterns of early and late sites.

While we are still focused on the detection of spatial clusters, the data format differs slightly from that in the previous chapter. Here we have point locations and two type of points: cases or controls (non-cases). In the Anasazi example, there is no obvious classification of early sites as “case” or “controls”, rather the choice is up to the analyst. The locations of early and late settlement sites appear in Figure 5.1 as well as locations of local rivers and streams. In the following sections, we apply ratio of density estimation and Generalized Additive Models (GAM) methods to this data set, then explore our spatial performance measures on them.

5.2 Methods

5.2.1 Density Ratio Method

The first analytic technique under consideration is due to Kelsall and Diggle who developed a non-parametric density ratio method based on kernel estimation (31; 32; 33). For a case/control study, we assume that the collections of case and control locations s follow heterogeneous Poisson processes with intensity functions $\eta_1(s)$ and $\eta_2(s)$, respectively. The log intensity ratio is defined by

$$\rho(s) = \log\{\eta_1(s)/\eta_2(s)\}.$$

When conditioning on the number of case and control sites n_1 and n_2, the data can be regarded as a pair of independent random samples with bivariate distributions across the study area with probability densities $f(s)$ and $g(s)$, where $f(s)$ and $g(s)$ are proportional to $\eta_1(s)$ and $\eta_2(s)$.

$$\begin{aligned}\rho(s) &= \log\frac{f(s)/\int_R \eta_1(s)ds}{g(s)/\int_R \eta_2(s)ds} \\ &= \log\{f(s)/g(s)\} - \log\frac{\int_R \eta_1(s)ds}{\int_R \eta_2(s)ds}\end{aligned}$$

Since the second part is a constant across the region R, the spatial variation in the log intensity ratio is proportional to the log density ratio, ie.,

$$\hat{\rho}(s) \propto \log\{\hat{f}(s)/\hat{g}(s)\}$$

where $\hat{f}(s)$ and $\hat{g}(s)$ are kernel estimators of $f(s)$ and $g(s)$ respectively.

In order to provide a measure of statistical significance, Kelsall and Diggle propose a

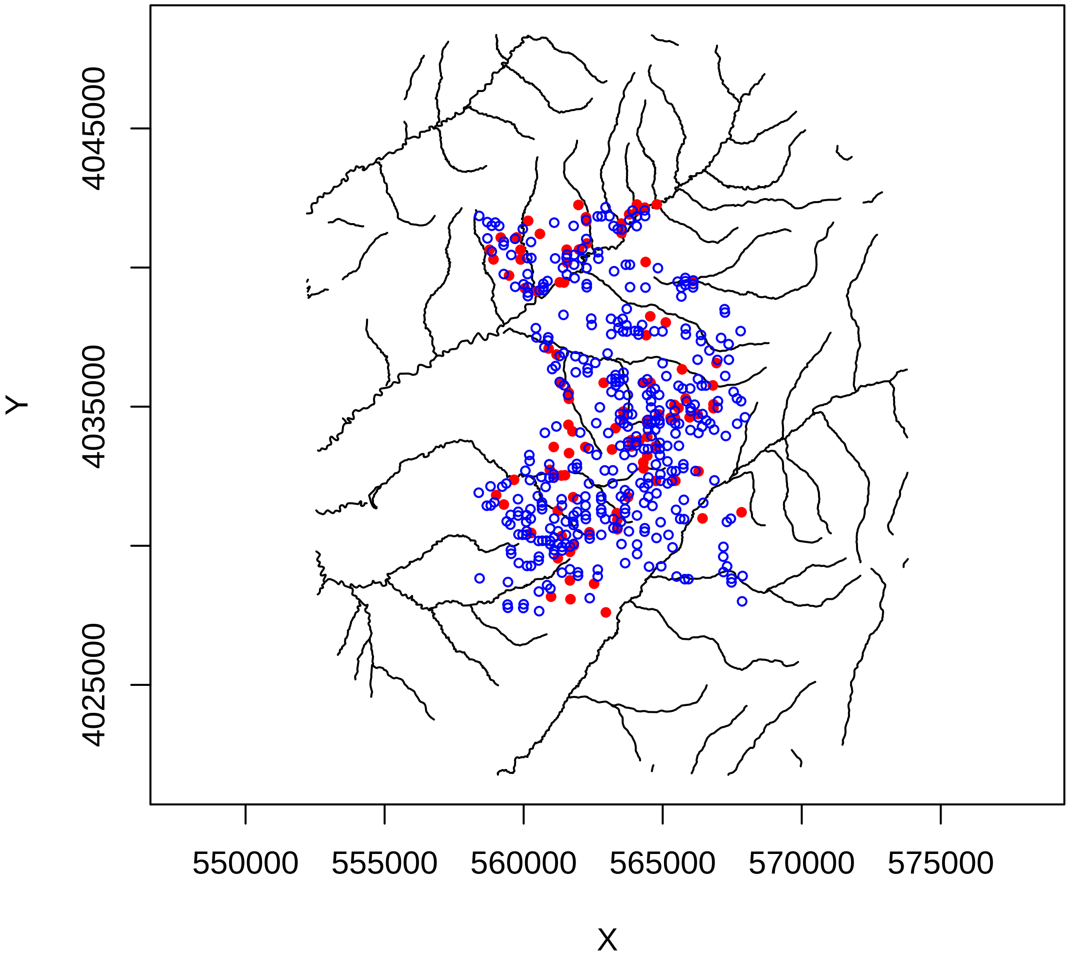


Figure 5.1: Settlement sites, red indicates early sites, and blue indicates later sites

tolerance interval for $\hat{\rho}(s)$ defined by Monto Carlo simulation under a null hypothesis of random labelling (33; 42).

The computation can be performed in R using kernel estimation functions available in the KernSmooth package (15; 51). For the studies below, we use a standard bivariate normal density for our spatial 2-dimensional data.

5.2.2 Generalized Additive Models (GAM)

We next consider extensions to the kernel smoothing approaching that allow local covariate effects. Hastie and Tibshirani introduced Generalized Additive Models (GAMs) in 1984 based on nonparametric regression or smoothing techniques (37; 76; 77). GAMs are extensions of generalized linear models. In generalized linear models, we have

$$g(\mu) = \alpha + \sum_{i=1}^{p} x_i \beta_i$$

where $\mu = E(Y)$, Y is distributed according to a member of the exponential family, or with known mean-variance relationship, $g(.)$ denotes a link function, α is the intercept, x_i is the ith independent variable, and β_i is the ith parameter coefficient.

In GAMs, some or all parametric terms in GLM are replaced by smooth functions, such as LOESS or cubic splines, but the additive form is still kept.

$$g(\mu) = \alpha + \sum_{i=1}^{q} z_i \theta_i + f_1(x_1) + f_2(x_2, x_3) + \dots$$

where $\mu = E(Y)$, Y is distributed according to a member of the exponential family, or with known mean-variance relationship, $g(.)$ denotes a link function, α is the intercept, z_i is the ith independent variable, and θ_i is the ith parameter coefficient, the $f(.)$s are smooth functions, and x_1, x_2, x_3 are independent variables in the smooth terms.

If we estimate coefficients in GAMs by likelihood maximization, this will often lead to a overfitted wiggly model. To control the model smoothness, a penalty term is introduced in our objective functions, therefore, we maximize a penalized likelihood for the model.

$$l_p(\beta) = l(\beta) - \frac{1}{2}\sum_j \lambda_j \beta^T S_j \beta, \tag{5.1}$$

where λ is the smoothing parameter, and $\lambda \in [0, \infty)$, and $\frac{1}{2}\beta^T S_j \beta$ is the penalty for the jth smoothing term. When $\lambda = 0$, we obtain an un-penalized regression spline, and when $\lambda \to \infty$ we obtain a straight line estimate for f. Choosing the smoothing parameter λ is crucial to model fitting. If λ is too small, the data will be undersmoothed, if λ is too large, the data will be oversmoothed.

As a result, GAMs are often fitted through a penalized iteratively re-weighted least squares method given the penalized matrix, and the optimal fit is obtained when it converges. GAMs can also be fitted by backfitting (77), which iteratively smooth partial residuals from the model.

In spatial applications, it makes more sense to use 2-dimensional smooth functions to define surfaces. Therefore, thin plate splines which are suitable to multiple predictor variables (74) are applied here. Thin plate splines also have the attractive property of being equivalent to best linear unbiased spatial predictions obtained via kriging (19; 54; 55). A thin plate smoothing spline for estimating a 2-dimensional data

$$y_i = g(x_1, x_2) + \epsilon_i$$

minimizes

$$\parallel y - f \parallel^2 + \lambda J_{22}(f)$$

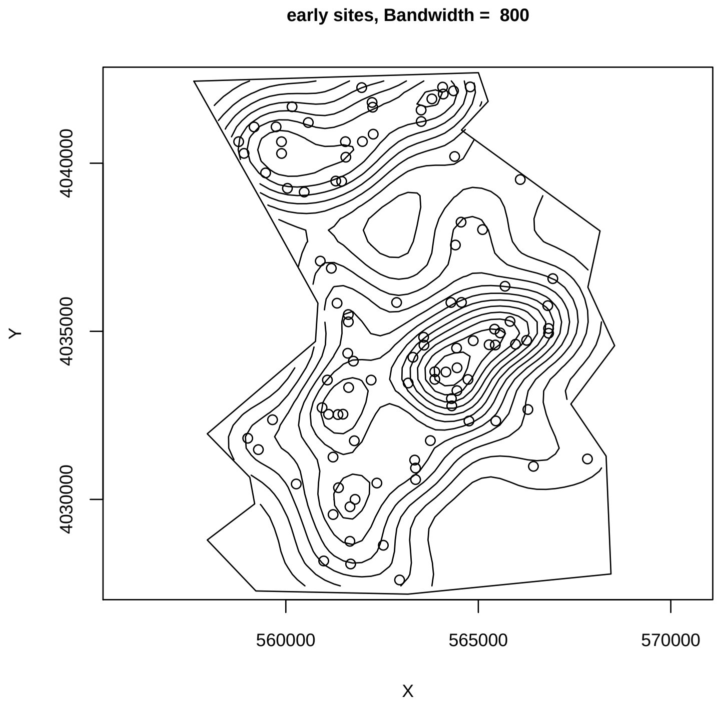
where λ is the smoothing parameter, and J_{22} is the penalty function for smoothing two predictor variables.

GAMs have been implemented in the "gam" and "mgcv" packages in R, providing a convenient computing platform for our studies.

5.3 Results

To begin, we present results based on applying Kelsall and Diggle's method to the observed Anasazi data. Figure 5.2 shows the estimated densities using Gaussian kernels with bandwidth 800 units for early and late sites. It reveals different patterns between early and late sites as shown in the log relative risk surface plot 5.3. Extreme high log relative risk appear at the right bottom corner as there is no early sites discovered in this region. The symbols "+" and "-" in the contour plot 5.4 indicate the areas where the estimated log relative risk surface is above or below the 95% pointwise tolerance intervals defined by 999 random labelling null hypothesis (i.e., randomly assignment of the "case" and "control" labels) .

The ratio of density estimation method reveals that neither early sites nor late sites are uniformly randomly distributed in the study area. The early sites tend to be clustered near the left middle near coordinates (560000,4035000), while late sites are clustered in the left and right lower corner as well as near coordinates (564000,4038000) when the bandwidth is set at 800 units. When the bandwidth is reduced to 400 units (Figure 5.5), the clusters of early site and late site are shown in Figure 5.6.



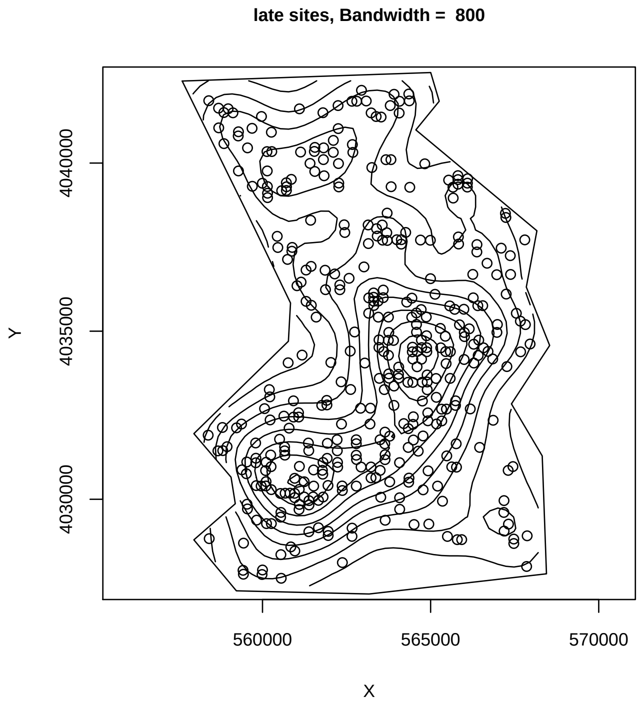


Figure 5.2: Estimated densities (normalized intensities) using Gaussian kernels and a bandwidth of 800 units for early and late sites in the Black Mesa

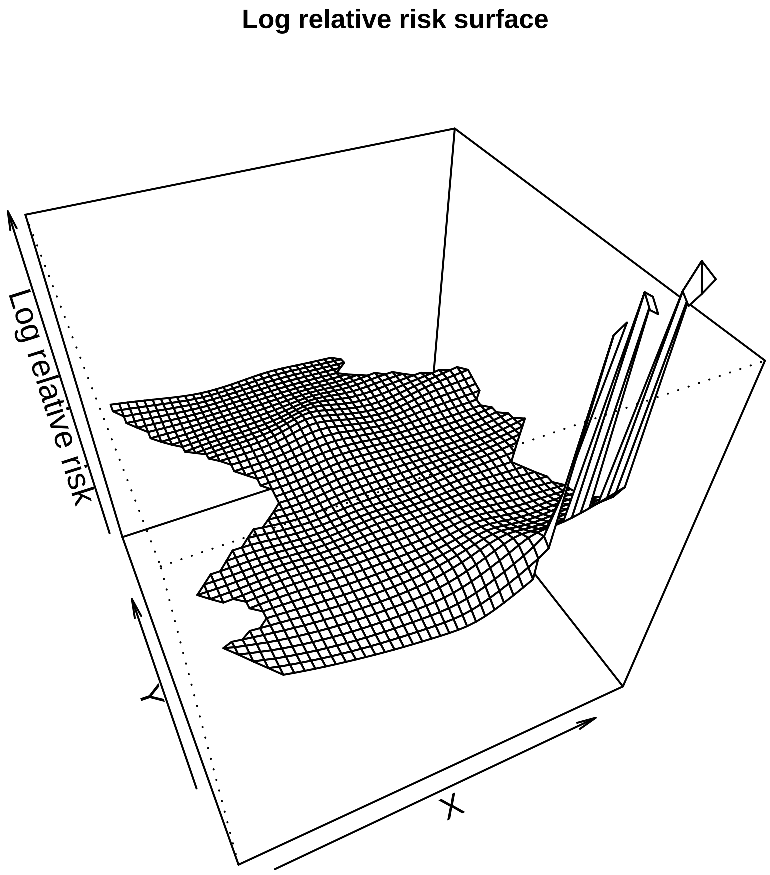


Figure 5.3: Estimated log relative risk surfaces in the Black Mesa Anasazi data

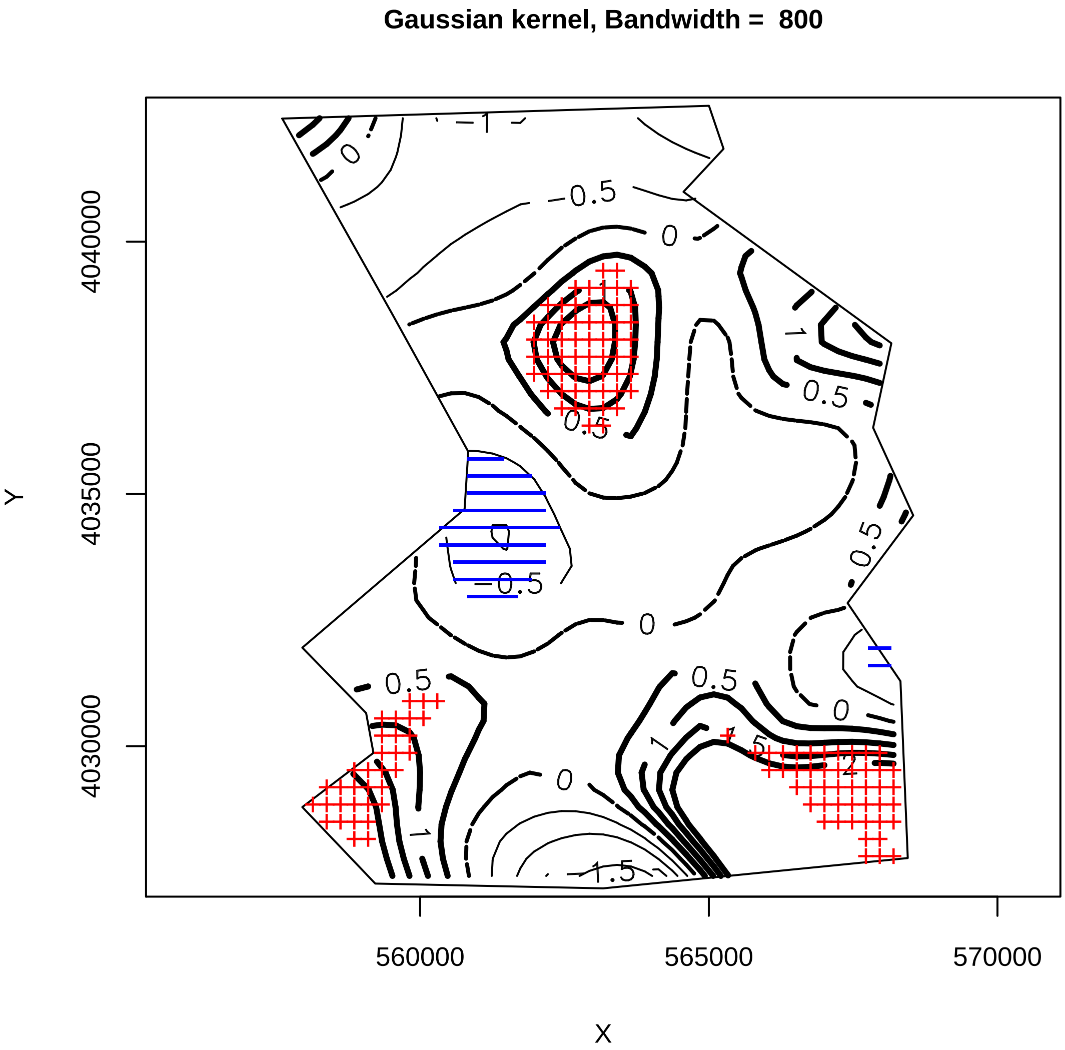


Figure 5.4: Contour plots in the Black Mesa Anasazi data

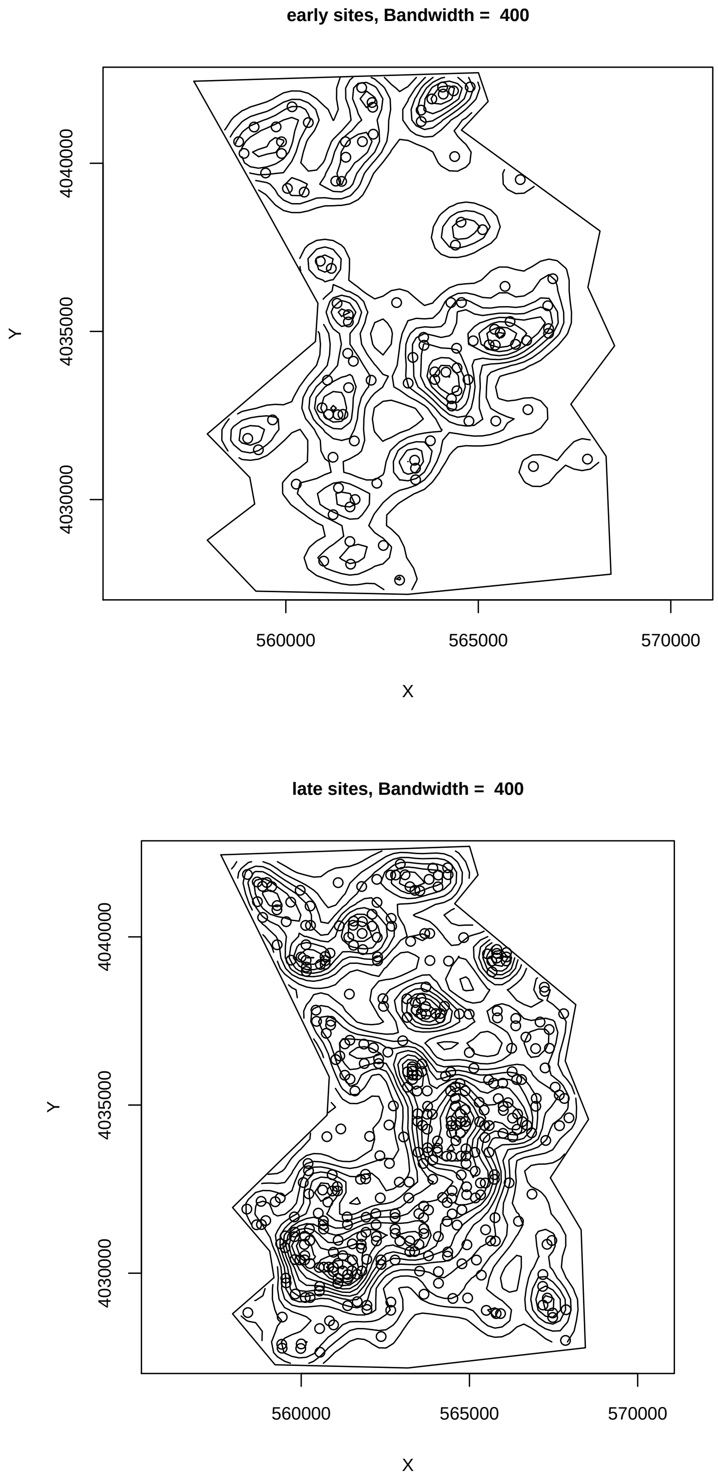


Figure 5.5: Estimated densities (normalized intensities) using Gaussian kernels and a bandwidth of 400 units for early and late sites in the Black Mesa Arizona data

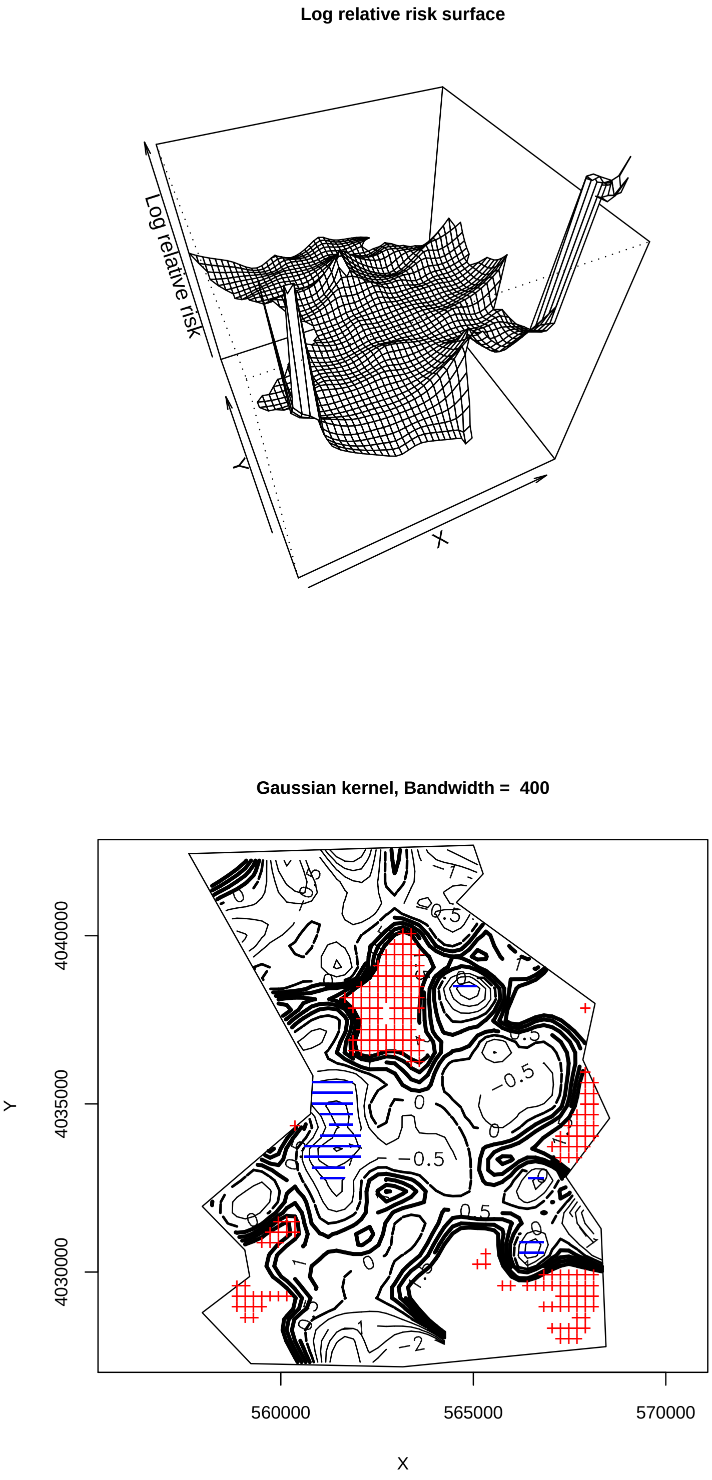


Figure 5.6: Estimated log relative risk surfaces (top) and contour plots (bottom) using Gaussian kernels and a bandwidth of 400 units for early and late sites in the Black Mesa Arizona data (top)

To explore the effect of streams around the sites, we next use GAMs and include the following covariates one at a time: minimum distance between a site to a stream, average distance between a site to streams within a 2,000 meter radius, number of streams within a 2,000 meter radius, total measure of segments of streams within 2,000 meters, average measure of segments of streams within 2,000 meters, and total measure of full streams who are within 2,000 meters of a site. The measure "full stream" is the length of the entire stream from end to end. In our results, we find that the natural log of the total measure of full streams is marginally significant. Therefore, if a site is close to longer streams, it is more likely to be a late site (Table 5.1). In addition, both models with or without covariates (Table 5.2) reveal that smooth terms are significant. Hence, spatial location associates with whether a site be early or late, above and beyond the impact of the observed local covariates.

Figure 5.7 shows the log relative risk based on GAM with or without covariates. Figure 5.8 show the contour plot, and the symbols "+" and "-" indicate the areas where the estimated log relative risk surface is above or below the 95% pointwise tolerance intervals defined by 999 random labelling null hypothesis.

Compared to results from kernel density estimation methods, generalized additive models showed that sites in the north region tend to be early sites. In addition, the areas with significant high or low relative risk of being an early site are generally larger than the results showed in previous section. Furthermore, the model including a covariate of total stream length reveals more detailed contours, and adapts the relative risk surface to the hydrographic landscape.

Table 5.1: GAM results with covariate

Parametric coefficients	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	-5.8609	3.7976	-1.543	0.123
Stream length	0.7492	0.3920	1.911	0.056
Approximate significance of smooth terms				
	edf	Est.rank	Chi.sq	p-value
s(X,Y)	10.79	22	41.28	0.00764

edf is estimated degrees of freedom.

Table 5.2: GAM results without covariate

Parametric coefficients	Estimate	Std. Error	z value	$Pr(>\|z\|)$
Intercept	1.4194	0.1168	12.16	$< 2e-16$
Approximate significance of smooth terms				
	edf	Est.rank	Chi.sq	p-value
s(X,Y)	13.32	27	47.24	0.00932

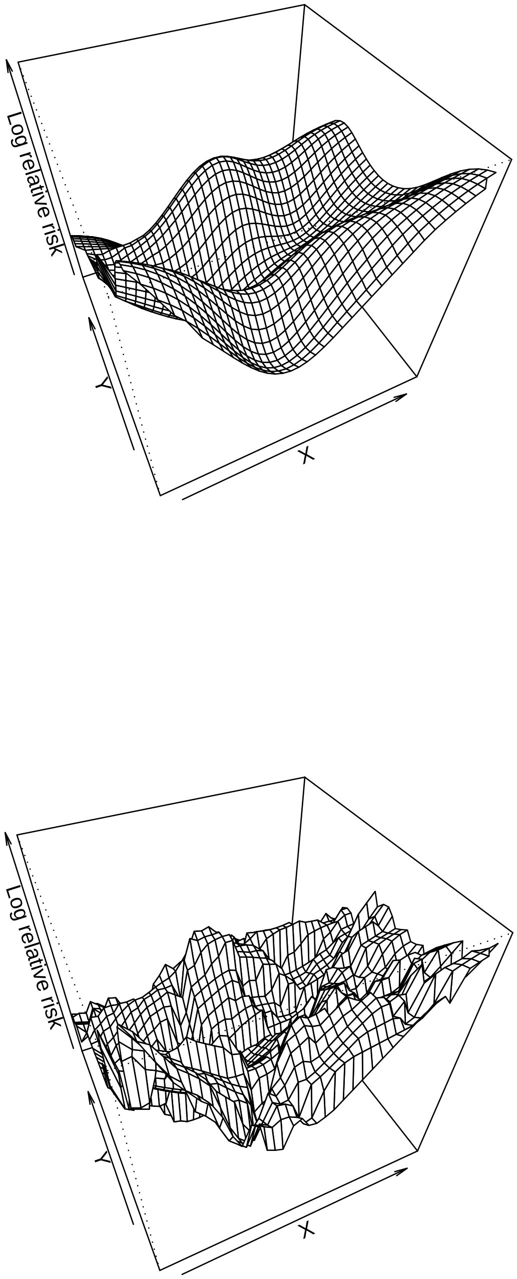


Figure 5.7: Estimated relative risk based on GAMs in the Black Mesa Anasazi data, the top panel shows the estimated log relative risk without covariate, and the bottom panel shows the estimated log relative risk with covariate.

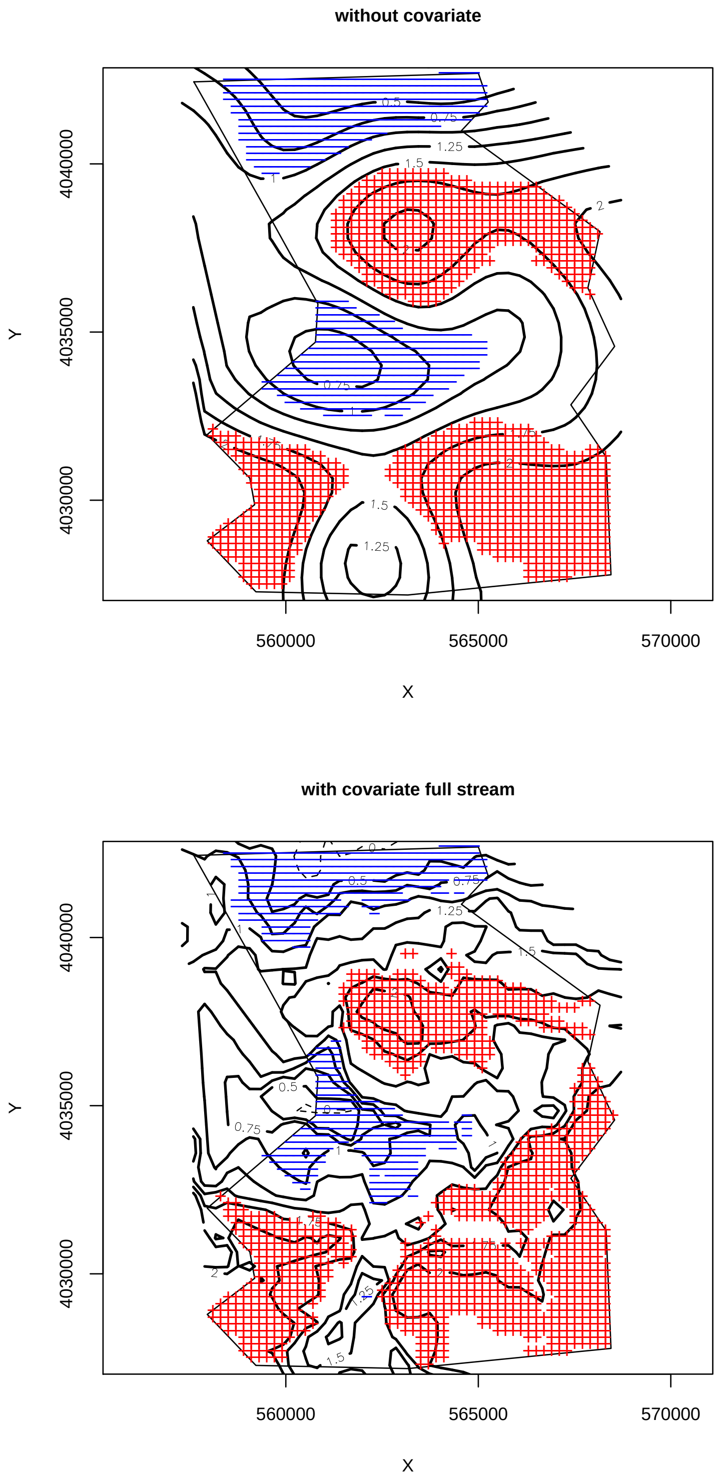


Figure 5.8: Estimated contour plots based on GAMs in the Black Mesa Anasazi data, the top panel shows the contour plot without covariate, and the bottom panel shows the contour plot with covariate.

5.4 Simulations and the Spatial AUC

In the previous chapter, we assessed spatial performance of count data. We apply similar ideas to regional point data in this chapter. Receiver operating characteristic (ROC) curves are popular summaries of statistical performance and combine information related to both sensitivity and specificity. In particular, the ROC curve plots sensitivity vs. false alarm rate (1- specificity), across a range of potential critical values. The area under the ROC curve (AUC) is a good indicator of the performance of a test. The closer the AUC is to 1.0, the better the test, and the closer the AUC is to 0.5, the worse the test. If the AUC is 1.0, the test is perfect. The ROC curve and AUC have seen broad application in many areas, including, but not limited to, medical decision making (53) and signal processing (72).

To evaluate statistical performance of spatial systems, we compute the ROC curve and AUC based on both kernel density ratio method and generalized additive models based on their ability to detect a hypothetical cluster representing a given increase in relative risk at a given location. By moving the location of this cluster around the study area, we can map the ability of each method to detect the cluster in any location. First, we simulate 1000 data sets based on the null hypothesis that the probability of an event being a case remains the same for all events. That is:

$$H_0 : \pi_i = \pi = \frac{\text{number of case}}{\text{total number of events}} \tag{5.2}$$

We also simulate data sets under the alternative hypothesis defined by a relative risk of 2.0 inside the cluster compared to outside the cluster, conditional on the numbers of case and control, i.e.,

$$H_a : RR = \frac{\pi_{in}}{\pi_{out}} = 2 \tag{5.3}$$

For the Anasazi data set with 100 early sites and 389 late sites, we set the cluster size to be 25. Inside the cluster, we randomly choose 10 sites to be early, and the other be late sites. Outside the cluster, we will randomly choose 90 sites to be early sites, and the rest will be late sites. Therefore, $\pi_{in} = 10/25 = 0.4$, and $\pi_{out} = \frac{90}{489-25} = 0.194$, and $RR = 0.4/0.194 = 2.06$, and a cluster is defined as a site and its 24 nearest neighbors. Figure 5.9 illustrate two such simulated data sets and identifies the clusters.

We use the same simulated data sets to compute the spatial AUC using both kernel density ratio method and generalized additive models. We next move the simulated cluster to different locations to see if performance changes with cluster location.

Based on the spatial AUC, the kernel density ratio method illustrates some areas with very good AUC values (above 0.9), but other areas with poor (nearly non-informative) AUC values ranging between 0.45 to 0.55, shown in Figure 5.10. The spatial AUC maps in Figure 5.11 for the generalized additive model approach shows that clusters in the corners are easier to identify and clusters in the middle are harder to identify. In addition, the addition of covariate information generates AUC performance above 0.75 for most areas.

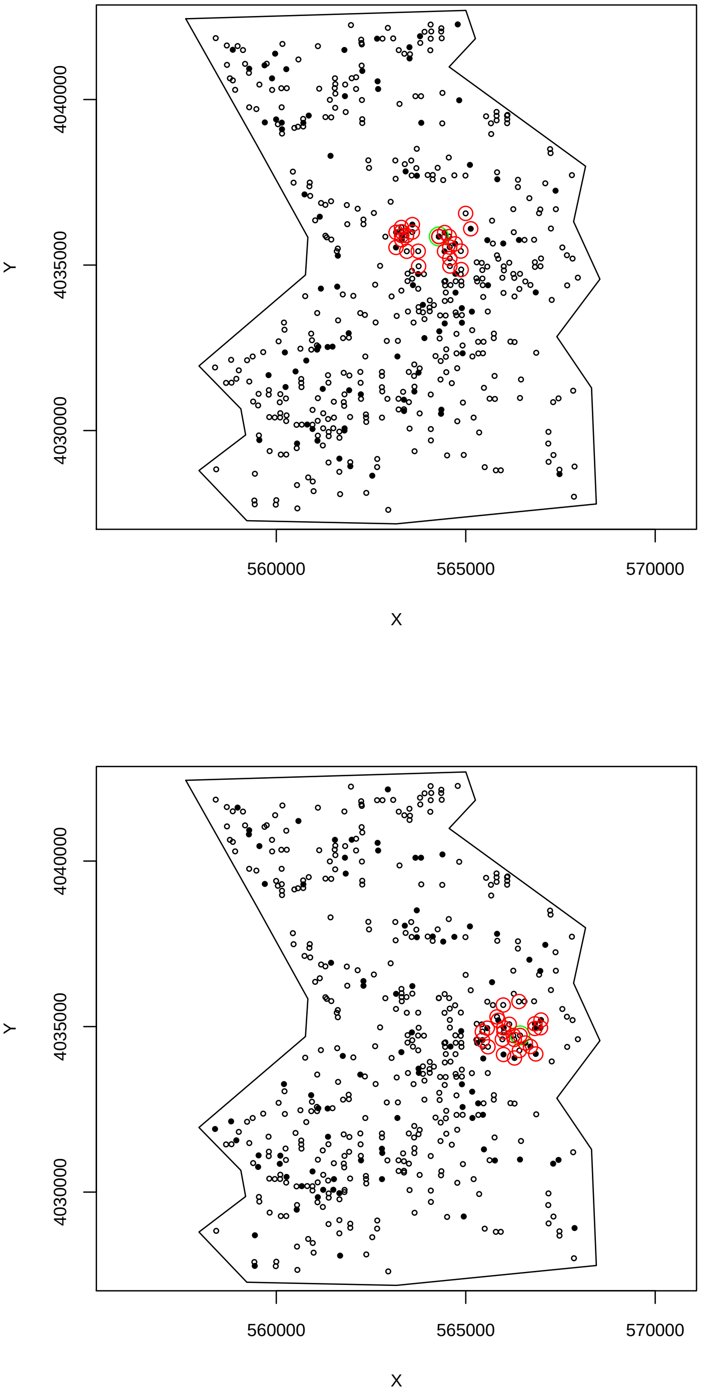
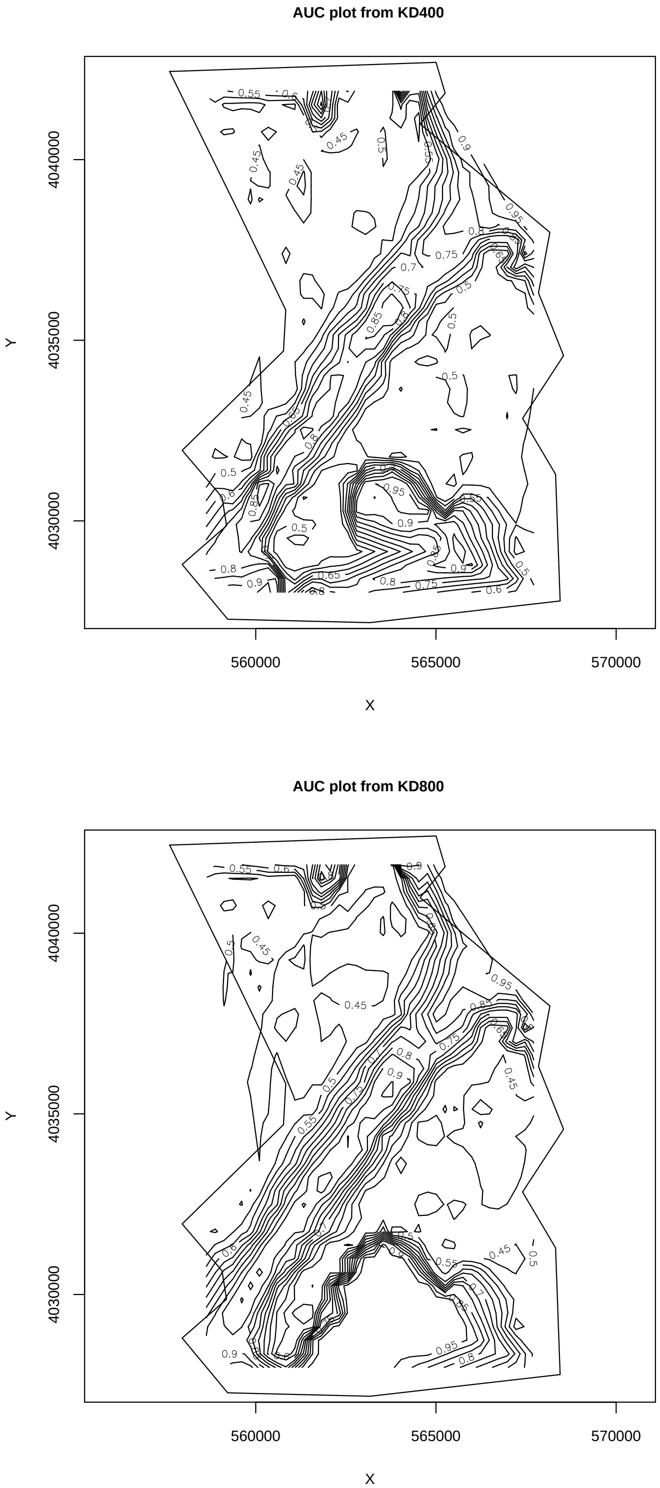


Figure 5.9: Simulated data sets with RR=2 inside the cluster. The green circle is the selected sites, the red circles indicated the 24 nearest neighbors. Open circles are late sites, closed circles are early sites



AUC plot from KD400

AUC plot from KD800

Figure 5.10: AUC plots based on KD method

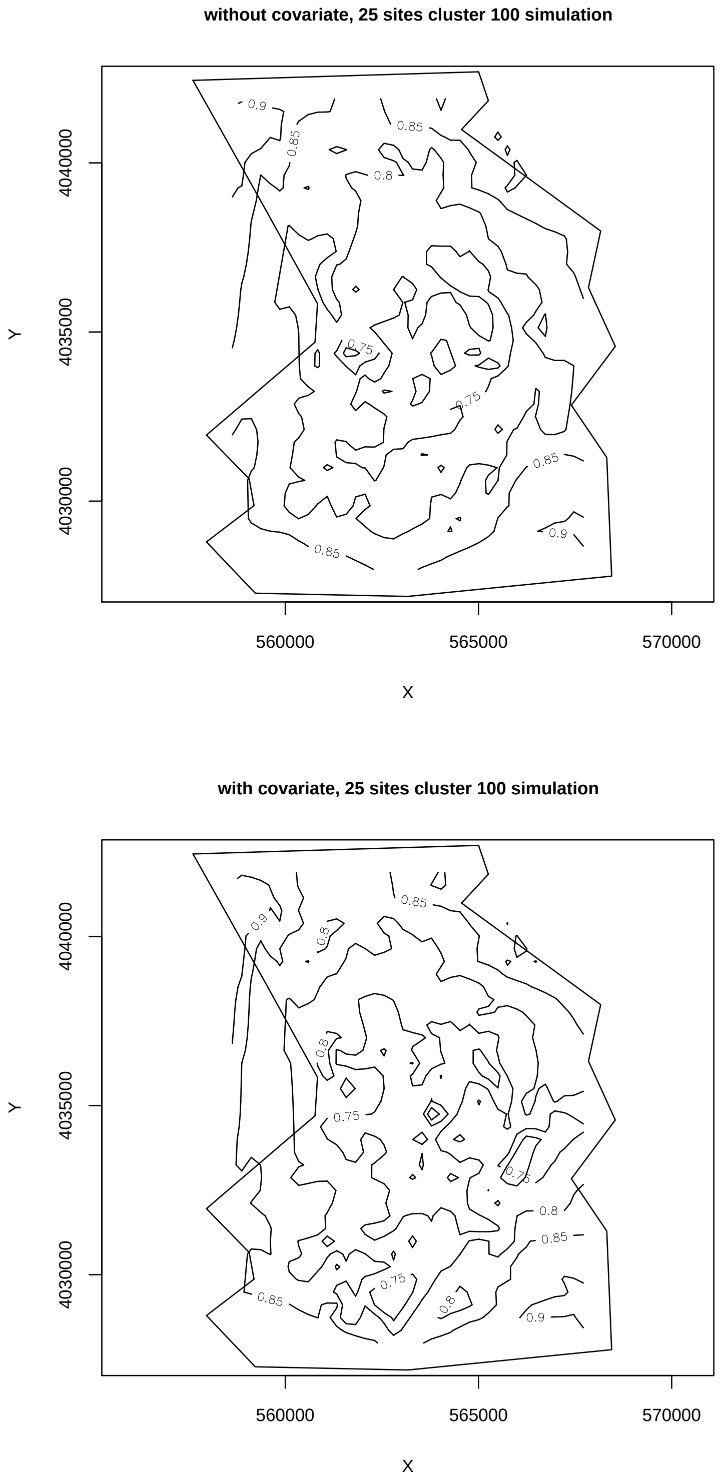


Figure 5.11: AUC plots based on GAM, the top panel shows AUC plot without covariates, and the bottom panel shows the AUC plot with covariates

5.5 Conclusions

In conclusion, we find that performance of cluster detection methods varies in space with the location of the cluster. The spatial AUC approach is effective for point data, and it can reveal the impact of data structure on cluster detection performance. In addition, local covariates also can influence detection and performance. For example, in the Black Mesa data set, late sites tend to be built near longer streams than are early sites. Finally, maps of performance can identify areas where clusters may be hard to detect.

Chapter 6

Spatial Performance for Outbreak Detection Systems

6.1 Introduction

In previous chapters we compared four different approaches for detecting outbreaks in surveillance data. In each case we essentially have several local tests investigating whether each local area (census tract or grid point) is part of a hot spot. In public health surveillance, detection of a cluster in a local area may affect larger areas by declaring a state of emergency or elevating security level for the entire country, even warning the whole world. For example, if one local clinic found evidence of human-to-human transmission of avian influenza, a global response would ensue. In this chapter we shift our focus from the detection of an individual cluster to the performance of these individual components as elements in an overall detection system.

To improve system performance, more than one test could be performed by each local office to detect an outbreak. For example, we may compute test statistics based on different data sources such as over-the-counter medicine sales or emergency room visits as in syndromic surveillance; we may also run different models on the same data. Based on results in previous chapters, we are particularly interested in combining tests with different levels of performance in such a way as to ensure acceptable global detection performance by the system. Here we explore decision theoretic approaches to combine multiple local tests into a system, and evaluating the statistical performance of such systems.

Some ways to combine local tests and maintain acceptable system level performance include Bonferroni adjustment, false discovery rates, and decision fusion. Bonferroni adjustments tend to increase the type II error while attempting to maintain Type I error rates (60; 78). The false discovery rate (FDR) is the expected proportion of false positive predictions among a large number of declared significant results. FDR has been widely used in the multiple comparisons problem in the fields of gene expression arrays, proteomics, and imaging analysis (22; 30), and FDR could be much higher than the traditional p-value critical value of 0.05.

A global system-wide decision could be based on all the information passed to and processed by a central office, and this is referred as data fusion. On the other hand, each local office could process the information first, and send its binary decision (outbreak/no outbreak) to the central office, then the central office could make a global decision based on the reports of these local offices, and this is referred as decision fusion. By this decision fusion approach, the central office does not require transmitting and storing huge amounts of information nor does it require the ability to process this information in a short time period, but some information may be lost. In this chapter, we focus on the statistical performance of the decision fusion approach.

Theory and application of data fusion rapidly developed during the last two decades especially in engineering and defense technology fields (6; 11). In particular, we follow work by Carol Lin (11) and add an explicitly spatial dimension to the problem. We focus on a decision fusion approach due to its direct links to system design and decision theory. In addition, the decision fusion approach takes into account disease prevalence information, as detailed in the following section. In contrast, it is hard to implement prevalence information directly within Bonferroni adjustment or false discovery rate methods.

6.2 Decision Fusion

Carol Lin et. al. evaluated the performance of systems with multiple tests and cost constraints (11). To define ideas, suppose we have two types of detectors: A and B. Further suppose that A has better sensitivity and specificity than B, but A costs more money. Given the total cost of the system, we are trying to design a system with better performance than either one of its components.

Let u_i denote the binary decision of detector i, i.e.,

$$u_i = \begin{cases} 0 & \text{detector } i \text{ decides } H_0 \\ 1 & \text{detector } i \text{ decides } H_1 \end{cases} \tag{6.1}$$

The optimal system level decision is based on decisions of individual detectors (5; 83), i.e.,

$$u = \gamma(u_1, ..., u_n) = \begin{cases} 0 & \text{fusion center decides } H_0 \\ 1 & \text{fusion center decides } H_1 \end{cases} \tag{6.2}$$

Let $J(u, H_j)$ be the cost of central decision choosing u (0 or 1) when H_j is true. To minimize this cost, the optimal decision of the central could be derived from the likelihood ratio test, i.e.,

$$\frac{P(u_1, ..., u_n|H_1)}{P(u_1, ..., u_n|H_0)} \gtrless_{H_0}^{H_1} \frac{P(H_0)[J(1, H_0) - J(0, H_0)]}{P(H_1)[J(0, H_1) - J(1, H_1)]} = \beta \tag{6.3}$$

If we assume a minimum probability of error criterion, $J(0,H_0) = J(1,H_1) = 0$, and $J(1,H_0) = J(0,H_1) = 1$, then the right side of equation (6.3) only depends on the prevalance of the disease. If we assume statistical independence among local detectors, the left side of equation (6.3) would be determined by the performance of individual tests A and B in the following manner,

$$\begin{aligned}
\frac{P(u_1,...,u_n|H_1)}{P(u_1,...,u_n|H_0)} &= \prod_{i=1}^{n} \frac{P(u_i|H_1)}{P(u_i|H_0)} \\
&= \prod_{S_1} \frac{P(u_i|H_1)}{P(u_i|H_0)} \prod_{S_0} \frac{P(u_i|H_1)}{P(u_i|H_0)} \\
&= \prod_{S_1} \frac{P_{D_i}}{P_{F_i}} \prod_{S_0} \frac{1-P_{D_i}}{1-P_{F_i}} \\
&= \prod_{S_1} \frac{P_{DA}}{P_{FA}} \prod_{S_1} \frac{P_{DB}}{P_{FB}} \prod_{S_0} \frac{1-P_{DA}}{1-P_{FA}} \prod_{S_0} \frac{1-P_{DB}}{1-P_{FB}} \\
&= \left(\frac{P_{DA}}{P_{FA}}\right)^{a_1} \left(\frac{P_{DB}}{P_{FB}}\right)^{b_1} \left(\frac{1-P_{DA}}{1-P_{FA}}\right)^{a_0} \left(\frac{1-P_{DB}}{1-P_{FB}}\right)^{b_0} \qquad (6.4)
\end{aligned}$$

where P_{D_i} indicates the probability of detection of the ith detector, P_{F_i} indicates the false alarm rate of the ith detector, P_{DA} and P_{DB} indicate the probability of detection of the A and B detectors respectively, P_{FA} and P_{FB} indicate the false alarm rate of A and B detectors respectively, S_1 indicates the set of detectors claiming positive detection, S_0 indicates the set of detectors claiming false detection, a_1 is the number of A detectors claiming positive, a_0 is the number of A detectors claiming negative, b_1 is the number of B detectors claiming H_1, and b_0 is the number of B detectors claiming H_0. Therefore, by taking the natural logarithm on both sides of equation (6.3) the decision rule could be represented as:

$$\begin{aligned} & a_1 \times \log\left(\frac{P_{DA}}{P_{FA}}\right) + b_1 \times \log\left(\frac{P_{DB}}{P_{FB}}\right) \\ & + a_0 \times \log\left(\frac{1-P_{DA}}{1-P_{FA}}\right) + b_0 \times \log\left(\frac{1-P_{DB}}{1-P_{FB}}\right) \gtrless_{H_0}^{H_1} \log(\beta) \end{aligned} \tag{6.5}$$

The performance of the system wide decision would be measured as:

$$\begin{aligned} P_F &= P(u=1|H_0) \\ &= P\left[a_1 \times \log\left(\frac{P_{DA}}{P_{FA}}\right) + b_1 \times \log\left(\frac{P_{DB}}{P_{FB}}\right)\right. \\ &\quad \left. + a_0 \times \log\left(\frac{1-P_{DA}}{1-P_{FA}}\right) + b_0 \times \log\left(\frac{1-P_{DB}}{1-P_{FB}}\right) > \log(\beta)|H_0\right] \end{aligned} \tag{6.6}$$

$$\begin{aligned} P_D &= P(u=1|H_1) \\ &= P\left[a_1 \times \log\left(\frac{P_{DA}}{P_{FA}}\right) + b_1 \times \log\left(\frac{P_{DB}}{P_{FB}}\right)\right. \\ &\quad \left. + a_0 \times \log\left(\frac{1-P_{DA}}{1-P_{FA}}\right) + b_0 \times \log\left(\frac{1-P_{DB}}{1-P_{FB}}\right) > \log(\beta)|H_1\right] \end{aligned} \tag{6.7}$$

Therefore, given the total number of A detectors (n_a), the number of A tests claiming an outbreak (a_1) follows a binomial distribution. Under the null hypothesis H_0,

$$P(n=a_1) = \binom{n_a}{a_1} P_{DA}^{a_1}(1-P_{DA})^{n_a-a_1}.$$

Under the alternative hypothesis H_1,

$$P(n = a_1) = \binom{n_a}{a_1} P_{FA}^{a_1}(1 - P_{FA})^{n_a - a_1}.$$

Similarly, the number of B tests claiming an outbreak (b_1) following a binomial distribution, and under the null hypothesis H_0,

$$P(n = b_1) = \binom{n_b}{b_1} P_{DB}^{b_1}(1 - P_{DB})^{n_b - b_1}.$$

Under the alternative hypothesis H_1,

$$P(n = b_1) = \binom{n_b}{b_1} P_{FB}^{b_1}(1 - P_{FB})^{n_b - b_1}.$$

Therefore, combining all the possible system outcomes that claim an outbreak, we obtain the system probability of detection with n_a A and n_b B detectors, denoted P_D.

$$P_D = \sum_{u=1} \binom{n_a}{a_1} P_{DA}^{a_1}(1 - P_{DA})^{n_a - a_1} \times \binom{n_b}{b_1} P_{DB}^{b_1}(1 - P_{DB})^{n_b - b_1} \tag{6.8}$$

Similarly, we could also obtain the system false alarm rate with n_a A and n_b B detectors, denoted P_F.

$$P_F = \sum_{u=1} \binom{n_a}{a_1} P_{FA}^{a_1}(1 - P_{FA})^{n_a - a_1} \times \binom{n_b}{b_1} P_{FB}^{b_1}(1 - P_{FB})^{n_b - b_1} \tag{6.9}$$

In Lin's thesis (11), she concluded that increasing the number of tests will improve the performance of the system in general, but that some counterintuitive results occur for systems comprised of small numbers of tests. Lin (11) also extended her results to assess the performance of collections of correlated tests.

6.3 Comprehensive Evaluation

To further explore the statistical performance for outbreak detection systems, we use Lin's general results to explore performance of outbreak detection systems with two types of detectors (A and B) with different cost and different sensitivity and specificity to find the best system under the constraint of total cost.

First, we consider the situation where the cost of the outbreak detection systems is C_t, and the cost of A and B are C_A and C_B respectively. Assume all the potential systems satisfy

$$n_A \times C_A + n_B \times C_B = C_t,$$

where n_A is the number of A detectors, and n_B is the number of B detectors. Furthermore, we assume the sensitivity of A and B are P_{DA} and P_{DB}, respectively, and the false alarm rates of A and B are P_{FA}, P_{FB}, respectively. The prevalence of the outbreak is $P(H_1) = 0.01$. In addition, we assume P_{DA}=0.80, P_{DB}=0.55, P_{FA}=0.20, P_{FB}=0.45, C_A=20, C_B=1, $P(H_0)$=0.99, $P(H_1)$=0.01. The cost ratio of A and B is 20:1. We consider total costs C_t of 100, 200 or 500. Therefore, n_A could range from 0 to $n_{A,max} = \lfloor C_t/C_A \rfloor$, and given the number of A detectors, the number of B detectors is defined, ie,

$$n_B = \lfloor (C_t - n_A \times C_A)/C_B \rfloor.$$

Hence, the best system can be chosen from these $n_{A,max}$+1 potential systems matching the cost constraint..

Next, we will relax the constraint of the fixed total cost, and consider only an upper bound, i.e.,

$$n_A \times C_A + n_B \times C_B \leq C_t.$$

In this case, given any number of A detectors, the number of B detectors could go from zero to the maximum that the system could afford. Therefore, the number of possible designs would be much larger than the previous one. Among these designs, we may also find a system with similar performance but lower cost than C_t.

6.4 Results

To begin, we consider systems with fixed total cost and use Lin's results to balance between the number of expensive, precise detectors and inexpensive, imprecise detectors. Figure 6.1 shows the performance of a set of systems composed of expensive, precise A detectors and inexpensive, imprecise B detectors. As noted above, A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The cost of the system is 100. These assumptions yield six systems as listed in Table 6.1. The left panel in Figure 6.1 demonstrates the probability of detection of the systems, and the right panel shows the false alarm rate of the systems. The same color points in these four panels correspond to the same system. For example, a system with zero expensive, precise detector and 100 inexpensive, imprecise detectors has a probability of detection of 0.0951, and false alarm rate of 0.0005; while a system with one expensive, precise detector and 80 inexpensive, imprecise detectors has a probability of detection of 0.1747, and false alarm rate of 0.0010.

In general, the probability of detection and false alarm rate of the system is increasing with increasing number of precise detectors, but decreasing with the increasing number of imprecise detectors. However, given our assumed performance values for individual detectors, a system with five expensive, precise detectors and no inexpensive, imprecise detector has a lower probability of detection, and lower false alarm rate than a system with four expensive, precise detectors and twenty inexpensive, imprecise detectors. To see why, note that, again under our assumptions, a system with only five expensive, precise detectors requires all five detectors declare positive in order for the system to declare positive that satisfy the following condition,

$$(a_1 - a_0) \times \log \frac{0.8}{0.2} > \log(\beta) \tag{6.10}$$

On the other hand, the system with four expensive, precise detectors, twenty inexpensive, imprecise detectors will need all four expensive and at least eight inexpensive or three expensive and at least fifteen inexpensive ones to declare positive for the system in order to satisfy the following condition,

$$(a_1 - a_0) \times \log \frac{0.8}{0.2} + (b_1 - b_0) \times \log \frac{0.55}{0.45} > \log(\beta) \tag{6.11}$$

Table 6.1: A set of systems of expensive, precise A detectors and inexpensive, imprecise B detectors. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 100.

Number of expensive, precise detectors	Number of inexpensive, imprecise detectors	Probability of detection of the system	False alarm rate of the system
0	100	0.0951	0.0005
1	80	0.1747	0.0010
2	60	0.2281	0.0011
3	40	0.3041	0.0012
4	20	0.4085	0.0014
5	0	0.3277	0.0003

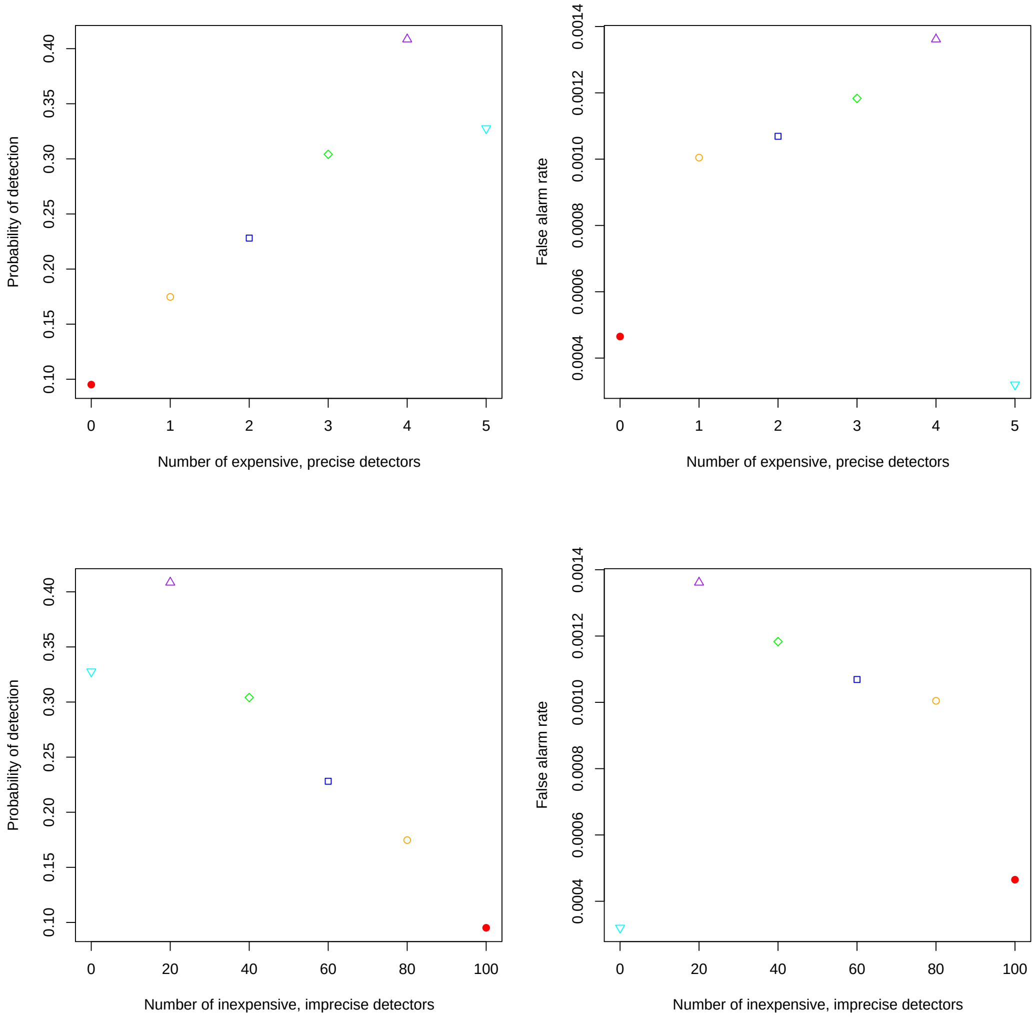


Figure 6.1: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 100. Each color represents one system.

If we double the allowable cost of the system to 200, the probability of detection of the system is dramatically improved compared to the system with total cost of 100 as shown in Figure 6.2. Table 6.2 shows the eleven ways to comprise the system. In general, the probability of detection of the system is increasing with increasing number of precise detectors, and the false alarm rate is decreasing with the increasing number of imprecise detectors. It is interesting to see that systems with odd number of expensive, precise detectors have a linear decreasing false alarm rate as the number of precise detectors increasing, and the systems with even number of precise detectors have a flatter trend as noted in Figure 6.2. Table 6.3 shows the combinations of a_1 and b_1 that will declare an outbreak for the system. A system with just imprecise detectors will require 56% of them showing positive to declare system wide positive conclusion, and systems with six precise detectors can declare positive with 91% of imprecise detectors showing positive, but systems with more than six precise detectors can never declare positive without any positive precise detectors. A system with nine precise detectors will require at least five of them showing positive to declare system-wide positive. For systems with fewer than nine precise detectors, they all need some imprecise detectors showing positive to declare system-wide positive.

When we allow the cost of the system could go to 500, the performance is even better as shown in Figure 6.3.

Table 6.2: A set of systems of expensive, precise A detectors and inexpensive, imprecise B detectors. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 200.

Number of expensive, precise detectors	Number of inexpensive, imprecise detectors	Probability of detection of the system	False alarm rate of the system
0	200	0.4165	0.0012
1	180	0.5129	0.0015
2	160	0.5500	0.0013
3	140	0.6116	0.0013
4	120	0.6591	0.0012
5	100	0.6976	0.0011
6	80	0.7425	0.0011
7	60	0.7652	0.0009
8	40	0.8059	0.0009
9	20	0.8108	0.0007
10	0	0.8791	0.0009

Table 6.3: Conditions that a set of systems declare positive. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 200.

Number of expensive, precise detectors	Number of inexpensive, imprecise detectors	Number of a_1 in the system	Minimum number of b_1 in the system
0	200	0	112
1	180	0	105
		1	98
2	160	0	99
		1	92
		2	85
3	140	0	92
		1	85
		2	78
		1	72
4	120	0	86
		1	79
		2	72
		3	65
		4	58
5	100	0	79
		1	72
		2	65
		3	58
		4	52
		5	45
6	80	0	73
		1	66
		2	59
		3	52
		4	45
		5	38
		6	31
7	60	1	60
		2	52
		3	45
		4	38
		5	32
		6	25
		7	18
8	40	3	39
		4	32
		5	25
		6	18
		7	11
		8	4
9	20	5	18
		6	12
		7	5
		8	0
		9	0
10	0	7	0
		8	0
		9	0
		10	0

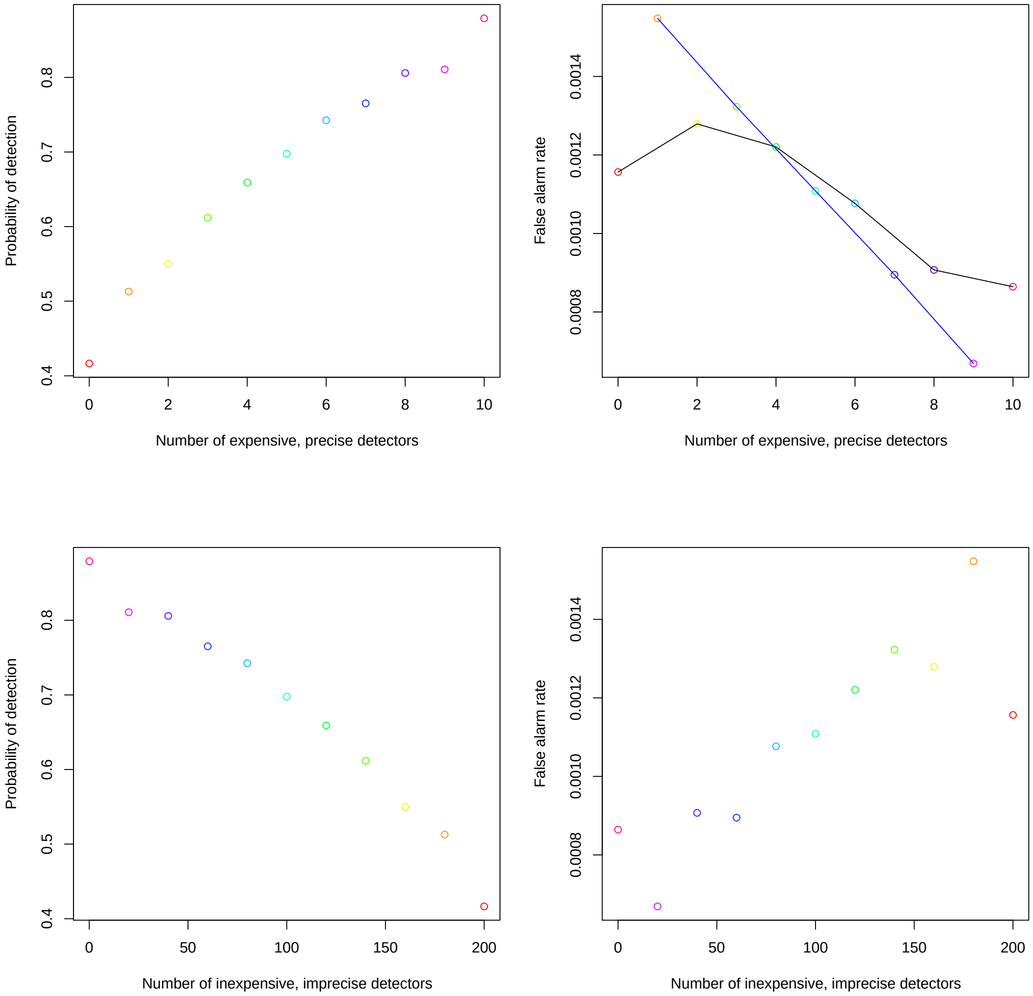


Figure 6.2: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 200. Each color represents one system. The connected lines in the upper right plot represent systems comprised of even (black) and odd (blue) numbers of detectors.

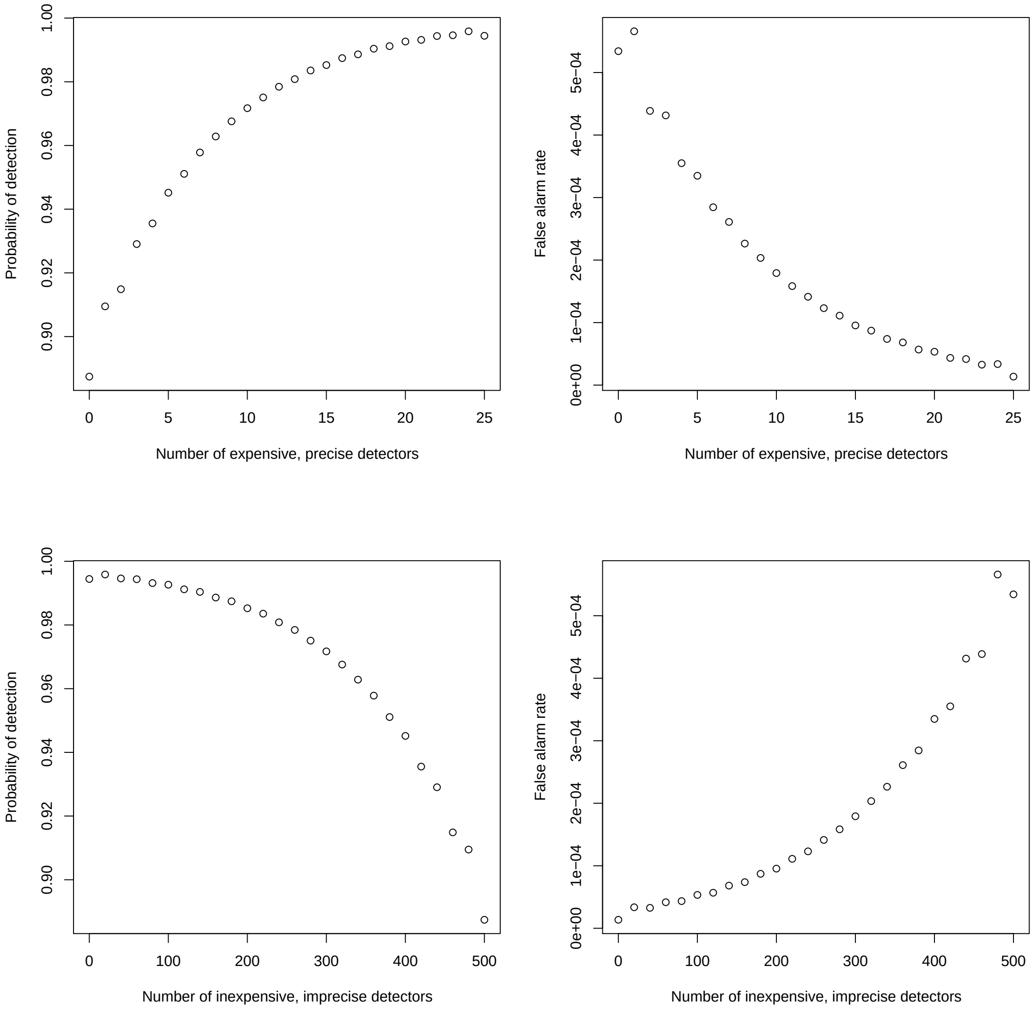


Figure 6.3: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The total cost of each system is 500.

Furthermore, if the performance of the imprecise detectors is slightly improved to a sensitivity of 0.6 and false alarm rate of 0.4, the performance of the systems is improved as shown in Figure 6.4, Figure 6.5 and Figure 6.6 for total cost of 100, 200 and 500 respectively. In contrast to the previous systems, in general, the probability of detection of the system is decreasing with increasing number of precise detectors, and the false alarm rate is increasing with the increasing number of imprecise detectors. Table 6.4 and Table 6.5 show six and eleven ways to comprise the systems of total cost of 100 and 200 respectively. Table 6.6 shows the combinations of a_1 and b_1 that will declare an outbreak for the system. A system with just imprecise detectors will require 53% of them showing positive to declare system wide positive, and systems with eight precise detectors can declare positive with 100% of imprecise detectors showing positive and no precise detector showing positive. A system with nine precise detectors can never declare positive with fewer than four positive precise detector. Hence, the influence of imprecise detectors to the system is more important when their performance improved. In addition, the systems with even numbers of expensive, precise detectors seem to have a distinct trend from the systems with odd number of expensive, precise detectors.

Table 6.4: A set of systems of expensive, precise A detectors and inexpensive, imprecise B detectors. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 100.

Number of expensive, precise detectors	Number of inexpensive, imprecise detectors	Probability of detection of the system	False alarm rate of the system
0	100	0.8211	0.00088
1	80	0.7879	0.00113
2	60	0.7112	0.00099
3	40	0.6762	0.00144
4	20	0.5646	0.00123
5	0	0.3277	0.00032

Table 6.5: A set of systems of expensive, precise A detectors and inexpensive, imprecise B detectors. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 200.

Number of expensive, precise detectors	Number of inexpensive, imprecise detectors	Probability of detection of the system	False alarm rate of the system
0	200	0.9812	0.00014
1	180	0.9775	0.00017
2	160	0.9697	0.00017
3	140	0.9661	0.00024
4	120	0.9549	0.00025
5	100	0.9487	0.00034
6	80	0.9331	0.00036
7	60	0.9226	0.00048
8	40	0.9003	0.00052
9	20	0.8815	0.00065
10	0	0.8791	0.00086

Table 6.6: Conditions that a set of systems declare positive. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 200.

Number of	Number of	Number of	Minimum number
expensive,	inexpensive,	a_1	of b_1
precise detectors	imprecise detectors	in the system	in the system
0	200	0	106
1	180	0	98
		1	94
2	160	0	90
		1	86
		2	83
3	140	0	81
		1	78
		2	74
		1	71
4	120	0	73
		1	70
		2	66
		3	63
		4	59
5	100	0	65
		1	61
		2	58
		3	54
		4	51
		5	48
6	80	0	56
		1	53
		2	50
		3	46
		4	43
		5	39
		6	36
7	60	0	48
		1	45
		2	41
		3	38
		4	34
		5	31
		6	28
		7	24
8	40	0	40
		1	36
		2	33
		3	30
		4	26
		5	23
		6	19
		7	16
		8	12
9	20	4	18
		5	14
		6	11
		7	8
		8	4
		9	1
10	0	7	0
		8	0
		9	0
		10	0

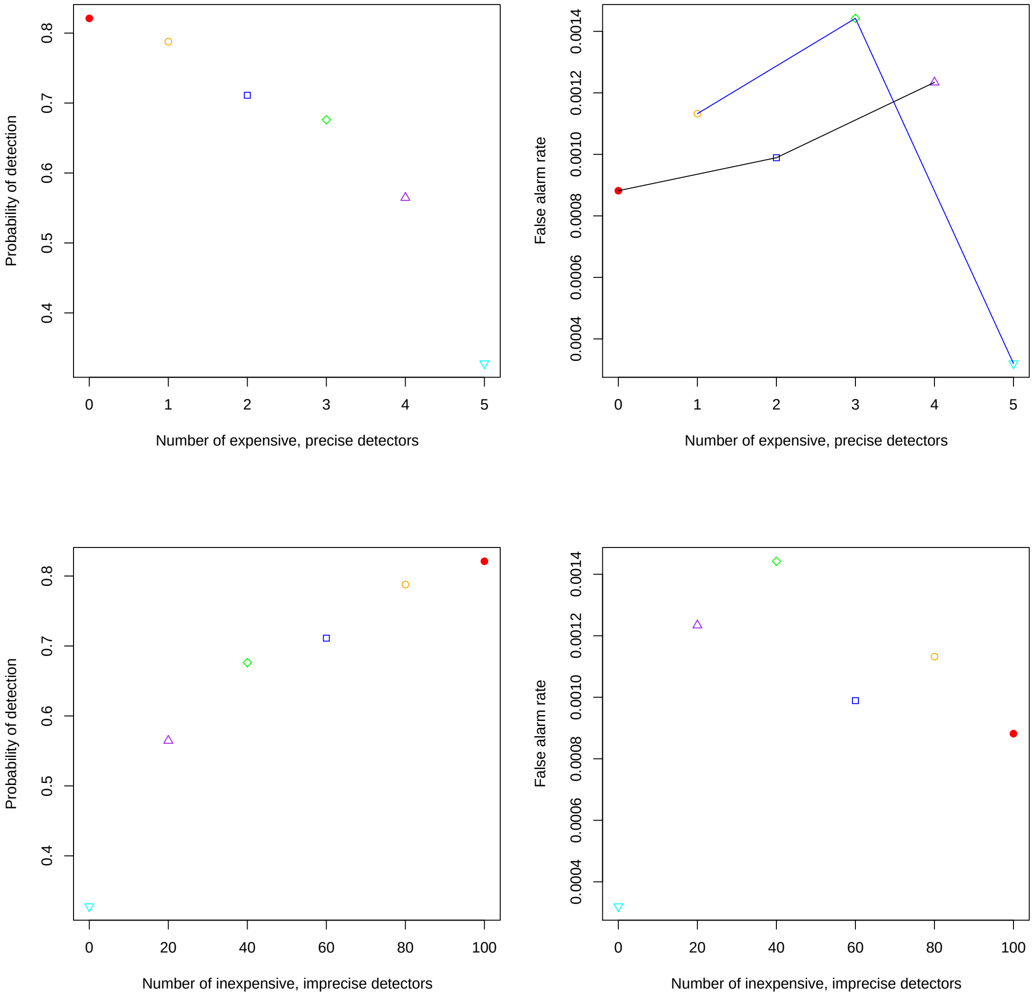


Figure 6.4: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 100. Each color represents one system. The connected lines in the upper right plot represent systems comprised of even (black) and odd (blue) numbers of detectors.

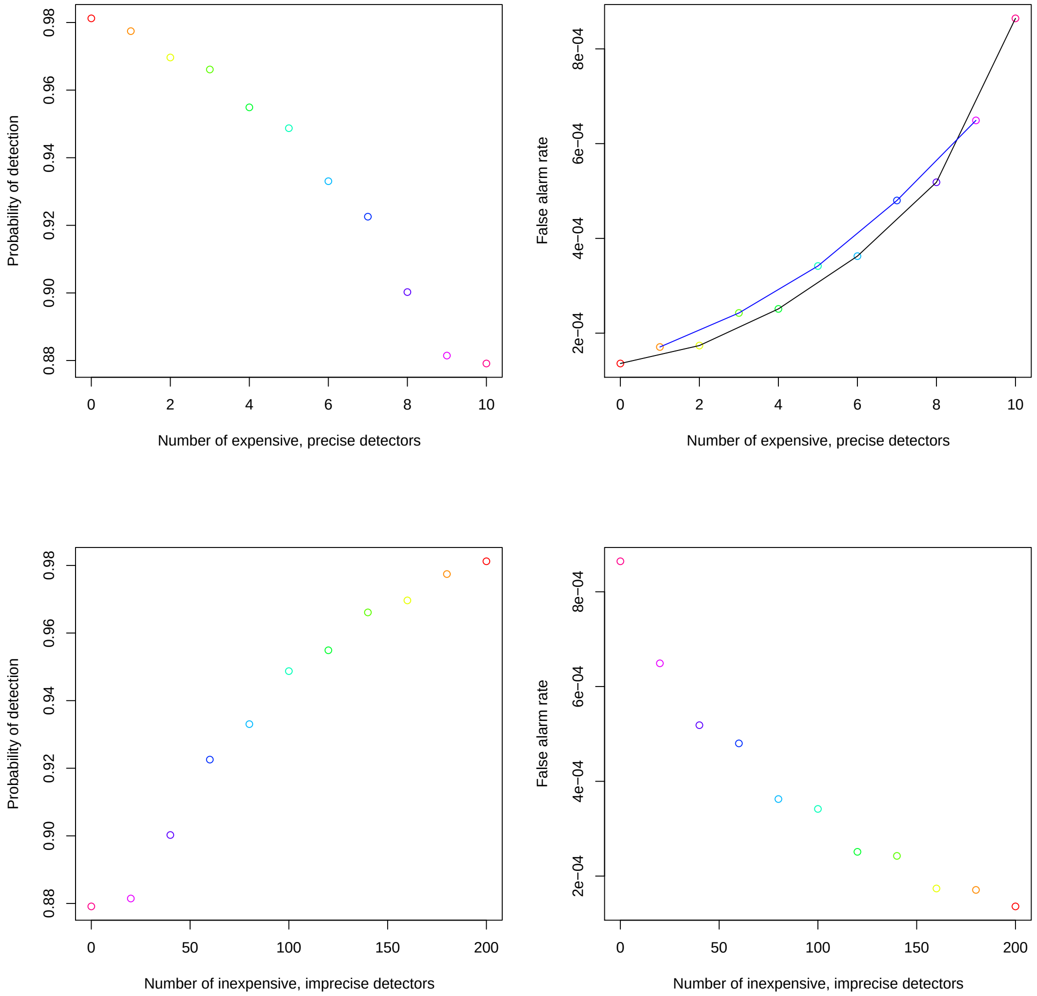


Figure 6.5: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 200. Each color represents one system. The connected lines in the upper right plot represent systems comprised of even (black) and odd (blue) numbers of detectors.

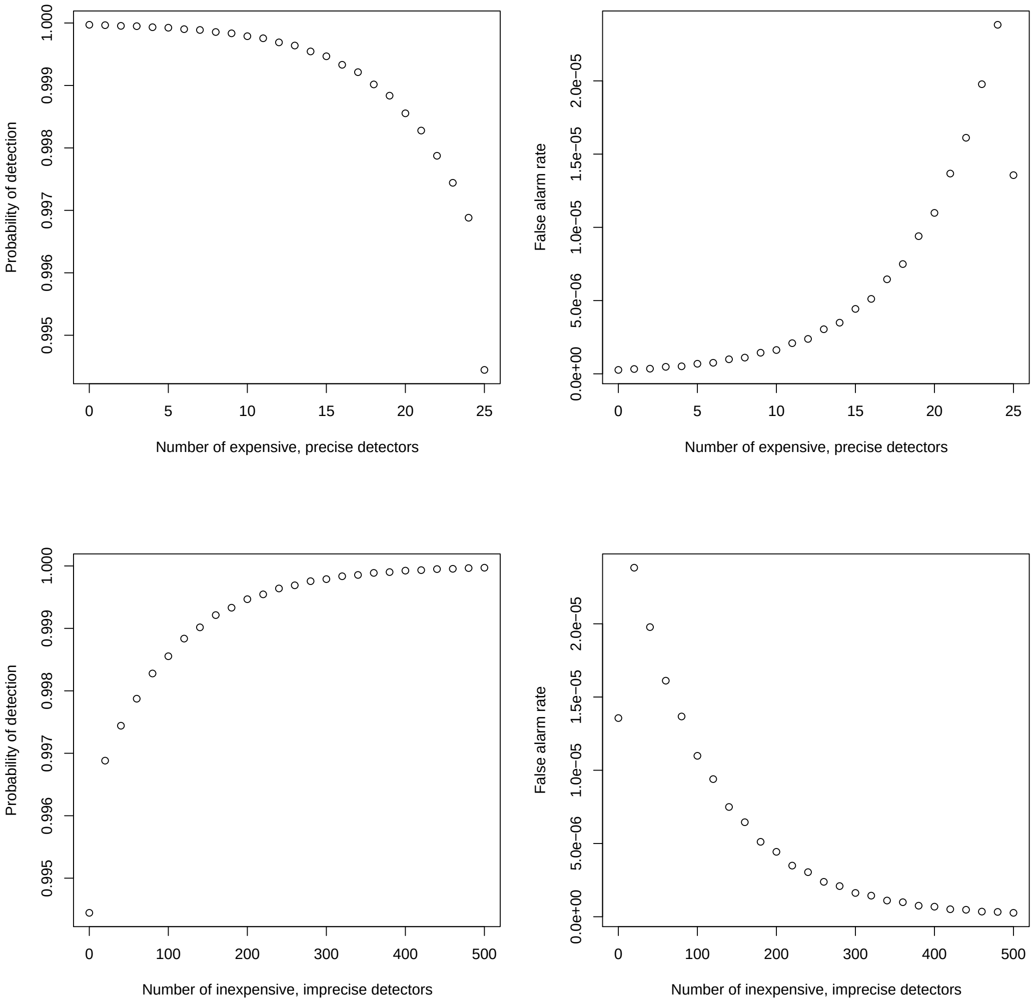


Figure 6.6: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The total cost of each system is 500.

When we consider Ct as the upper bound, we have many more possible systems to work with. For example, for the same resource available as in Table 6.1, a system could have three expensive detectors, and zero, one, two, ..., 40 imprecise detectors. Therefore, there are 306 possible systems given the same resource. The performance of these 306 systems are shown in Figure 6.7, which include systems showed in Figure 6.1 as marked by stars. Systems with one or two expensive, precise detectors have the maximum of probability of detection with maximum numbers of inexpensive detectors. Systems with zero, three or four precise detectors could use less than the maximum numbers of the imprecise detectors to reach better probability of detection. Except for systems comprised with four expensive detectors, adding more imprecise detectors increases the probability of detection, but also increases the false alarm rate as we expected. For every fixed number of precise detectors, there are two curves based on the number of imprecise detectors, one corresponds to an even number of imprecise detectors, the other for an odd number of imprecise detectors. This is because the number of detectors declaring outbreak required to trigger the system-wide warning follows different patterns for odd and even number of imprecise detectors. In general, systems with more precise detectors have higher probability of detection.

However, when the imprecise detector is slightly improved, systems with fewer precise detectors but many imprecise detectors have higher probability of detection than systems with more precise detectors as shown in Figure 6.8. Almost all the systems with the maximum number of imprecise detectors have better probability of detection than those with fewer than the maximum number of imprecise detectors given the same number of expensive detectors. Furthermore, the false alarm rate of the system increases when adding more imprecise detectors first, then reaches maximum and decreases afterwards.

Figure 6.9 and Figure 6.10 show the performance of systems with total cost of 200 with different imprecise detectors. The probability of detection of the system benefits from the increasing number of imprecise detectors, but false alarm rates are more complicated. Overall, systems with more precise detectors have higher probability of detection

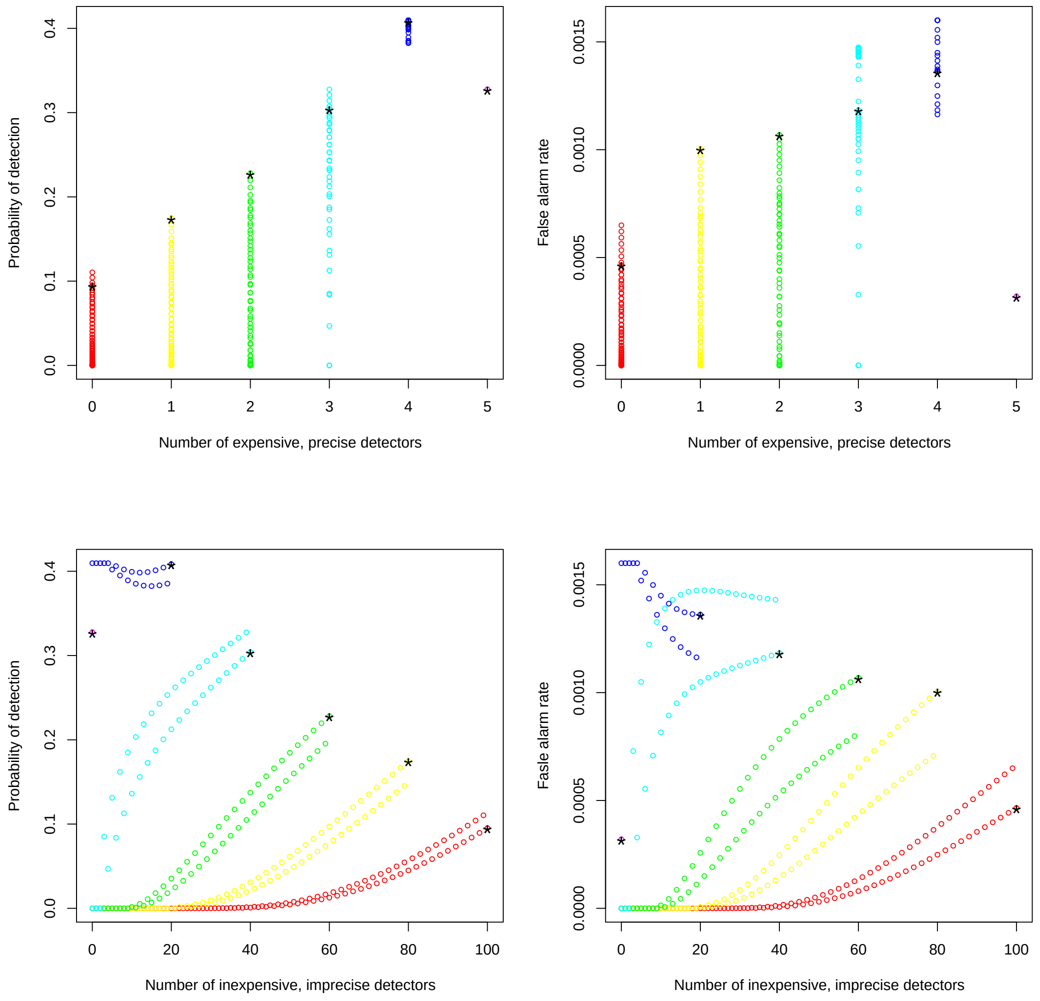


Figure 6.7: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The maximum cost of the system is 100. Systems with total cost of 100 are marked by symbol *, which are the systems showing in Figure 6.1. Each color represents systems with the same number of expensive, precise detectors.

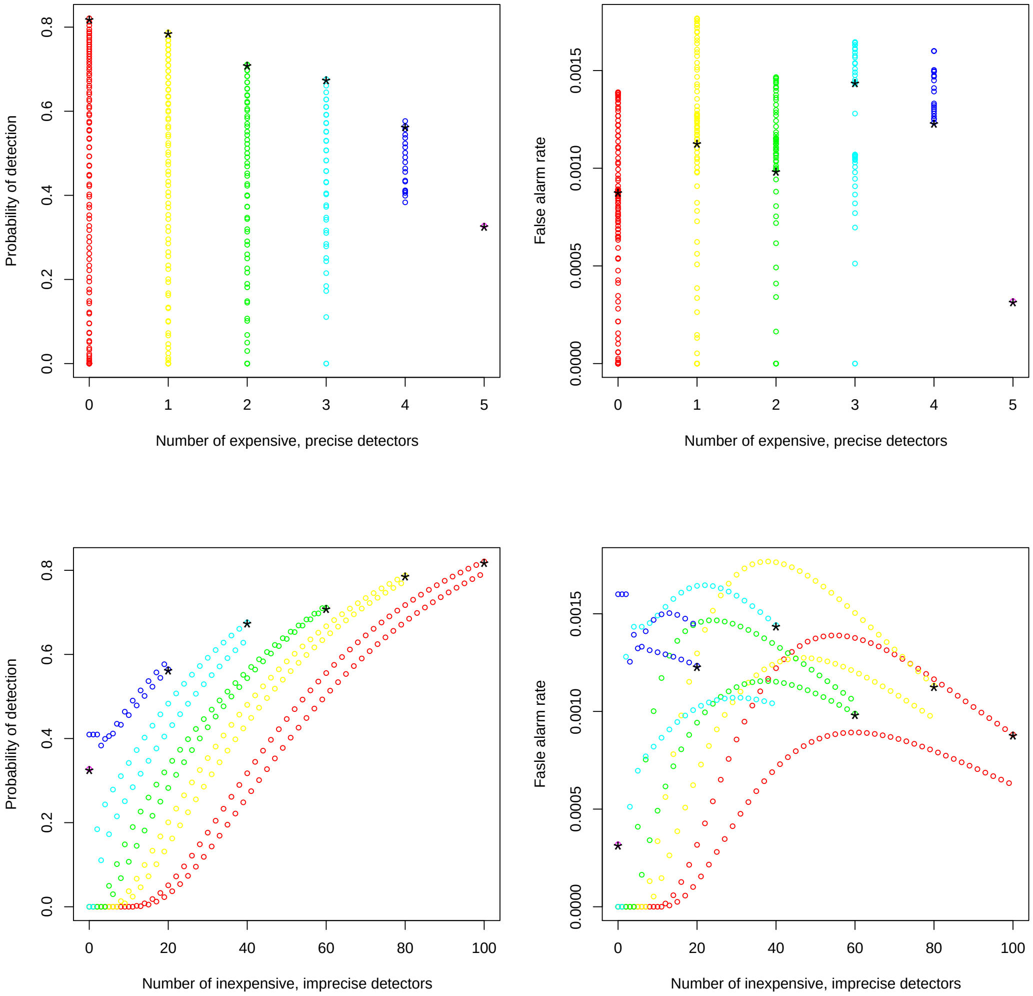


Figure 6.8: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The maximum cost of the system is 100. Systems with total cost of 100 are marked by symbol *, which are the systems showing in Figure 6.4. Each color represents systems with the same number of expensive, precise detectors.

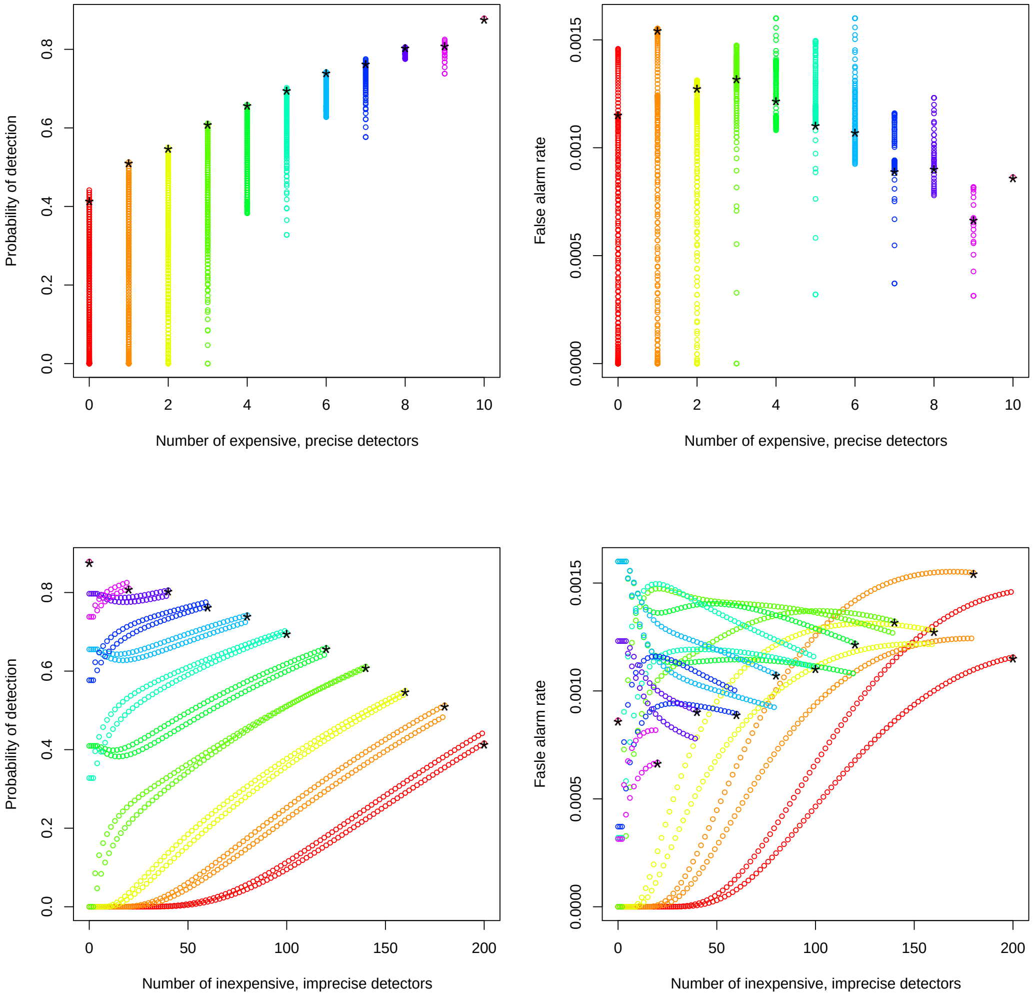


Figure 6.9: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.55 and false alarm rate of 0.45. The cost ratio between A and B is 20:1. The maximum cost of the system is 200. Systems with total cost of 200 are marked by symbol *, which are the systems showing in Figure 6.2. Each color represents systems with the same number of expensive, precise detectors.

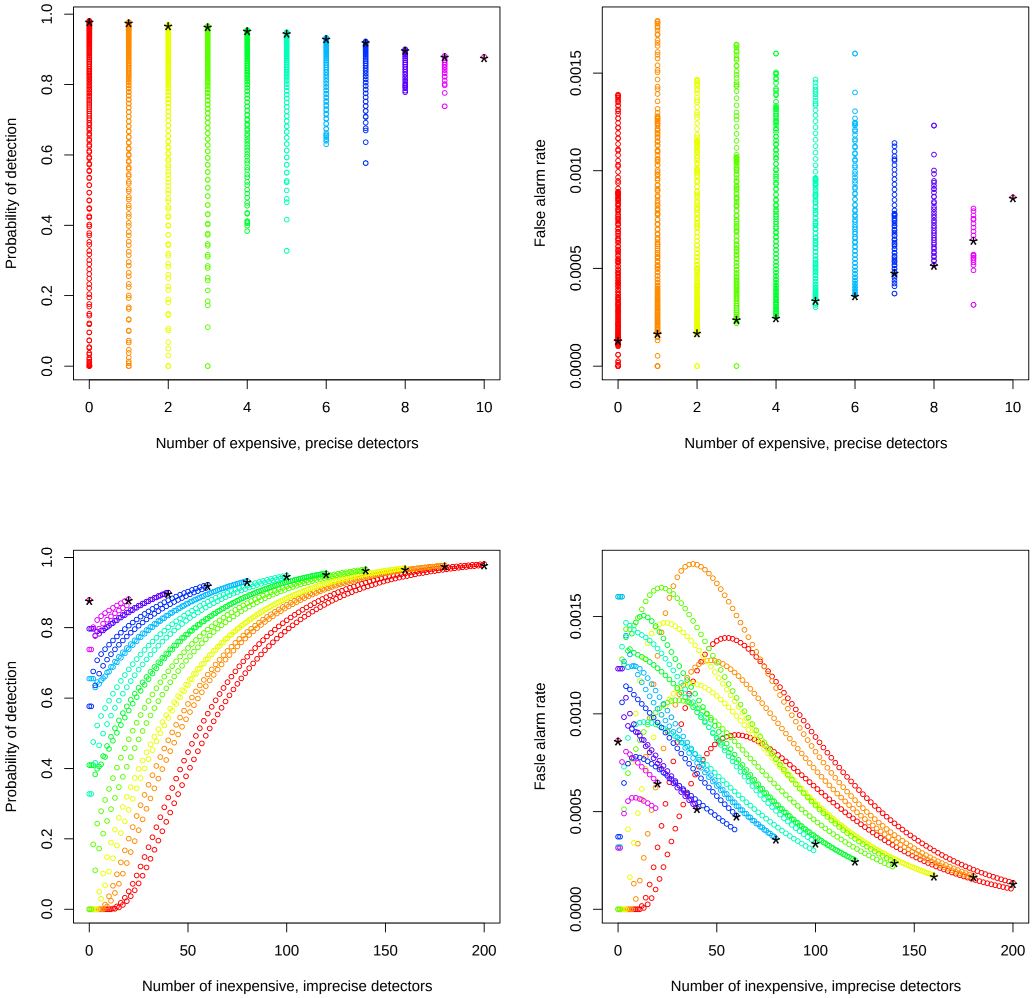


Figure 6.10: Performance of a set of systems with two types of detectors, A and B. A has sensitivity of 0.80 and false alarm rate of 0.2; while B has sensitivity of 0.6 and false alarm rate of 0.4. The cost ratio between A and B is 20:1. The maximum cost of the system is 200. Systems with total cost of 200 are marked by symbol *, which are the systems showing in Figure 6.5. Each color represents systems with the same number of expensive, precise detectors.

6.5 Conclusions

In summary, the performance of a system depends not only on the total allowable cost for the system, but also the performance of each individual detector, as well as the balance between expensive, precise and inexpensive, imprecise detectors. We quantify how, if we improve the performance of imprecise detectors even slightly, the performance of the resulting system improves dramatically, and patterns of performance can change. In addition, we show that lower-cost systems can perform as well as or better than systems expending the full allowable cost. These results indicate the need for careful calculation and computation to identify optimal system design, especially for systems comprised of small numbers of components.

Chapter 7

Summary and Future Research

In summary, we define and apply an approach using area under the ROC curve to evaluate statistical performance of spatial systems. We assessed spatial statistical performance of the spatial scan statistics SaTScan, Upper Level Set, and their applications of Santa Clara cardiac defects data. We defined how to assess spatial statistical performance of these methods and compare their results. The results showed that the SaTScan method performs better if the cluster is compacted, and the Upper Level Set Scan method is improving when the cluster is irregular shaped. The results also show that the Upper Level Set Scan Statistic can be quite sensitive to the presence of zero counts for very rare outcomes.

We also investigate the problem of cluster detection in regional point data, and compared spatial performance between Kernel Density method and Generalized Additive Models. We apply both methods to the Black Mesa archaeological project to identify clusters of early versus late Anasazi settlement sites when adjusting for exposure to rivers around those sites. We evaluate the spatial power of these applications, and generate the receiver operating characteristic (ROC) curve based on Monte Carlo simulations.

Finally, we assess performance of detection systems composite with multiple detectors with different sensitivity and false alarm rates based on decision fusion theory. In summary, the performance of a system depends not only on the total resources available, but also the performance of each individual detector, as well as the balance between expensive, precise and inexpensive, imprecise detectors. Slightly improving the inexpensive, imprecise detectors can dramatically increase the probability of detection of the system. Careful calculation or computation are required to identify an optimal system since it's hard to guess a good design. Non-intuitive results often appear for systems comprised of a small number of detectors.

In the future, we could explore other spatial detection methods. There are more than 100 global clustering methods reviewed by Kulldorff (46). We could also extend the system with more than two types of detectors, and investigate how each type of detectors would affect the performance of the system. In addition, we could extend these applications beyond public health, such as inspiring network sensor detection designs in signal precessing etc (66; 81). In addition, one could use other measures of performance in spatial settings, such as positive predictive values and negative predictive values.

Bibliography

[1] Goldenberg A, Shmueli G, Caruana RA, and Fienberg SE., *Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales*, Proceedings of the National Academy of Sciences of the United States of America **99** (2002), no. 8, 5237 –5240.

[2] Webb A, *Statistical pattern recognition*, John Wiley & Sons, 2002.

[3] Lawson AB and Kleinman K, *Spatial and syndromic surveillance for public health*, John Wiley & Sons, 2005.

[4] Thomas AJ and Carlin BP, *Late detection of breast and colorectal cancer in minnesota counties: an application of spatial smoothing and clustering*, Statistics in Medicine **22** (2003), 113–127.

[5] Reibman AR and Nolte LW, *Optimal detection and performance of distributed sensor systems*, IEEE Trans. Aerospace and Electronic Systems **AES-23** (1987), no. 1, 24–30.

[6] Dasarathy BV, *Decision fusion*, IEEE computer society press, Los Alamitos, CA, 1994.

[7] Green C, Hoppa RD, Young TK, and Blanchard JF, *Geographic analysis of diabetes prevalence in an urban area*, Social Science and Medicine **57** (2003), 551–560.

[8] Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, and Elliott P, *Statistical issues in the analysis of disease mapping data*, Statistics in Medicine **19** (2000), no. 17-18, 2493–2519.

[9] Witham CS and Oppenheimer C, *Mortality in england during the 17834 laki craters eruption*, Bulletin of Volcanology **67** (2004), 15–25.

[10] Trumbo CW, *Public requests for cancer cluster investigations: A survey of state health departments*, American Journal of Public Health **90** (2000), no. 8, 1300–1302.

[11] Lin CY, *Evaluating and optimizing performance of diagnostic systems with multiple tests*, Thesis, Emory University, 2005.

[12] Wartenberg D, *Investigating disease clusters: why, when and how?*, Journal of the Royal Statistical Society. Series A- Statistics in Society **164** (2001), 13–22.

[13] Martin DL, Goodman AH, Armelagos GJ, and Magennis AL, *Black mesa anasazi health: reconstructing life from patterns of death and disease*, Southern Illinois University Press, Carbondale, 1991.

[14] Bravata DM, McDonald KM, Smith WM, Rydzak C, Szeto H, Buckeridge DL, Haberland C, and Owens DK, *Systematic review: Surveillance systems for early dection of bioterrorism-related diseases*, Annals of Internal Medicine **140** (2004), no. 11, 910–922.

[15] Ruppert DR, Wand MP, and Carroll RJ, *Semiparametric regression*, Cambridge University, New York, 2003.

[16] Hill EG, Ding L, and Waller LA, *A comparison of three tests to detect general clustering of a rare disease in santa clara county, california*, Statistics in Medicine **19** (2000), no. 10, 1363–1378.

[17] Angulo F, Voetsch A, Vugia D, Hadler J, Farley M, Hedberg C, Cieslak P, Morse D, Dwyer D, Swerdlow D, and FoodNet Working group, *Determing the burden of human illness from foodborne diseases: Cdc's emerging infectious disease program foodborne disease active surveillance network (foodnet)*, Veterinary Clinics of North America: Food Animal Practice **14** (1998), 165–172.

[18] Mostashari F, Kulldorff M, Hartman JJ, Miller JR, and Kulasekera V, *Dead bird clustering: A potential early warning system for west nile virus activity*, Emerging Infectious Diseases **9** (2003), 641–646.

[19] Wahba G, *Comment on cressie*, The American Statistician **44** (1990), no. 3, 255–256.

[20] Patil GP and Taillie C, *Geographic and network surveillance via scan statistics for critical area detection*, Statistical Science **18** (2003), no. 4, 457–465.

[21] ______, *Upper level set scan statistic for detecting arbitrarily shaped hotspots*, Environmental and Ecological Statistics **11** (2004), no. 2, 183–197.

[22] Keselman HJ, Cribbie R, and Holland B, *Controlling the rate of type i error over a large set of statistical tests*, Britsh Journal of Mathematical & Statistical Psychology **55** (2002), 27–39.

[23] Burkom HS, *Development, adaptation, and assessment of alerting algorithms for biosurveillance*, Johns Hopkins APL technical digest **24** (2003), no. 4, 335 –342.

[24] Marschner IC and Bosch RJ, *Flexible assessment of trends in age-specific hiv incidence using two-dimensional penalized likekihood*, Statistics in Medicine **17** (1998), 1017–1031.

[25] Besag J and Newell J, *The detection of clusters in rare diseases*, Journal of the Royal Statistical Society. Series A - Statistics in Society **154** (1991), 143–155.

[26] Conley J, Gahegan M, and Macgill J, *A genetic approach to detecting clusters in point data sets*, Geographical Analysis **37** (2005), no. 3, 286–314.

[27] Glaz J and Balakrishnan N, *Scan statistics and applications*, Birkhauser, Boston, 1999.

[28] Mayer JD, *The role of spatial analysis and geographic data in the detection of disease causation*, Social Science & Medicine **17** (1983), no. 16, 1213–1221.

[29] Nordin JD, Goodman MJ, Kulldorff M, Ritzwoller DP, Abrams AM, Kleinman K, Levitt MJ, Donahue J, and Platt R, *Simulated anthrax attacks and syndromic surveillance*, Emerging Infectious Diseases **11** (2005), 1394 –1398.

[30] Storey JD and Tibshirani R, *Statistical significance for genomewide studies*, Proc. Natl Acad. Sci. USA **100** (2003), 9440–9445.

[31] Kelsall JE and Diggle PJ, *Kernel estimation of relative risk*, Bernoulli **1**, no. 1/2.

[32] ______, *Spatial variation in risk of disease: a nonparametric binary regression approach*, Applied Statistics **47**, no. 4.

[33] ______, *Non-parametric estimation of spatial variation in relative risk*, Statistics in Medicine **14** (1995), 2335–2342.

[34] Naus JI, *Clustering of random points in two dimensions*, Biometrika **52** (1965), no. 1/2, 263–267.

[35] ______, *The distribution of the size of the maximum cluster of points on a line*, Journal of the American Statistical Association **60** (1965), no. 310, 532–538.

[36] Brody JJ, *The anasazi*, Rizzoli International Publications INC., New York, NY, 1990.

[37] Chambers JM and Hastie TJ, *Statistical models in s*, Chapman & Hall, London, 1991.

[38] Jennings JM, Curriero FC, Celentano D, and Ellen JM., *Geographic identification of high gonorrhea transmission areas in baltimore, maryland*, American Journal of Epidemiology **161** (2005), 73–80.

[39] Kleinman K, Abrams A, Kulldorff M, and Platt R, *A model-adjusted space-time scan statistic with an application to syndromic surveillance*, Epidemiology and Infection **133** (2005), 409 –419.

[40] Kleinman K, Abrams A, Yih WK, Platt R, and Kulldorff M, *Evaluating spatial surveillance: detection of known outbreaks in real data*, Statistics in Medicine **25** (2006), no. 5, 755 –769.

[41] Rothman KJ, *A sobering start for the cluster busters conference*, American Journal of Epidemiology **132** (1990), no. 1, S6–S13.

[42] Waller LA, Hill EG, and Rudd RA, *The geography of power: Statistical performance of tests of clusters and clustering in heterogeneous populations*, Statistics in Medicine **25** (2006), no. 5, 853–865.

[43] Deane M, Swan SH, Harris JA, Epstein DM, and Neutra RR, *Adverse pregnancy outcomes in relation to water contamination, santa-clara county, california, 1980-1981*, American Journal of Epidemiology **129** (1989), no. 5, 894–904.

[44] Hills M and Alexander F, *Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: A review*, Journal of the Royal Statistical Society. Series A (Statistics in Society) **152** (1989), no. 3, 353–384.

[45] Kulldorff M, *A spatial scan statistic*, Communications in Statistics-Theory and Methods **26** (1997), no. 6, 1481–1496.

[46] ______, *Tests of spatial randomness adjusted for an inhomogeneity: A general framework*, Journal of the Amercian Statistical Association **101** (2006), no. 475, 1289–1305.

[47] Kulldorff M, Feuer EJ, Miller BA, and Freedman LS, *Breast cancer clusters in the northeast united states: A geographic analysis*, American Journal of Epidemiology **146** (1997), no. 2, 161–170.

[48] Kulldorff M, Huang L, Pickle L, and Duczmal L, *An elliptic spatial scan statistic*, Statistics in Medicine **25** (2006), no. 22, 3929–3943.

[49] Kulldorff M, Heffernan R, Hartman J, Assuncao R, and Mostashari F, *A space-time permutation scan statistic for disease outbreak detection*, PLOS MEDICINE **2** (2005), no. 3, 216–224.

[50] Lewis MD, Pavlin JA, Mansfield JL, O'Brien S, Boomsma LG, Elbert Y, and Kelley PW, *Disease outbreak detection system using syndromic data in the greater washington dc area*, American JOURNAL OF preventive Medicine **23** (2002), no. 3, 180–186.

[51] Wand MP and Jones MC, *Kernel smoothing*, Chapman and Hall, London, 1995.

[52] Clements MS, Armstrong BK, and Moolgavkar SH, *Lung cancer rate predictions using generalized additive models*, Biostatistics **6**, no. 4.

[53] Pepe MS, *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, New York, 2003.

[54] Cressie N, *Geostatistics*, The American Statistician **43** (1989), no. 4, 197–202.

[55] ______, *Reply*, The American Statistician **44** (1990), no. 3, 256–258.

[56] Boyle P, d'onofrio A, Maisonneuve P, Steveri G, Robertson C, Tubiana M, and veronesi U, *Measuring progress against cancer in europe: has the 15 % decline targeted for 2000 come about?*, Annals of Oncology **14** (2003), 1312–1325.

[57] Elliott P and Wartenberg D, *Spatial epidemiology: Current approaches and future challenges*, Environmental Health Perspectives **112** (2004), no. 9, 998–1006.

[58] Elliott P and Wakefield J, *Disease clusters: Should they be investigated, and, if so, when and how ?*, Journal of the Royal Statistical Society. Series A- Statistics in Society **164** (2001), 3–12.

[59] Sebastiani P, Mandl KD, Szolovits P, Kohane IS, and Ramoni MF, *A bayesian dynamic model for influenza surveillance*, Statistics in Medicine **25** (2006), no. 11, 1823 –1825.

[60] Bender R and Lange S, *Adjusting for multiple testing - when and how?*, Journal of clinical epidemiology **54** (2001), no. 4, 343–349.

[61] Brookmeyer R and Stroup DF, *Monitoring the health of populations, statistical principles & methods for public health surveillance*, Oxford University Press, New York, 2004.

[62] Chen R, Connelly RR, and Mantel N, *Analyzing post-alarm data in a monitoring-system in order to accept or reject the alarm*, Statistics in Medicine **12** (1993), no. 19-20, 1807–1812.

[63] Heffernan R, Mostashari F, Das D, Kulldorff M, and Weiss D, *Syndromic surveillance in public health practice, new york city*, Emerging Infectious Diseases **10** (2004), 858–864.

[64] Marshall RJ, *A review of methods for the statical-analysis of spatial patterns of*

disease, Journal of the Royal Statistical Society. Series A. General **154** (1991), 421–441.

[65] Neutra RR, *Counterpoint from a cluster buster*, American Journal of Epidemiology **132** (1990), no. 1, 1–8.

[66] Barbarossa S and Scutari G, *Bio-inspired sensor network design*, IEEE Signal Processing Magazine **24** (2007), no. 3, 26–35.

[67] Cousens S, Everington D, Ward HJT, Huillard J, Will RG, and Smith PG, *The geographical distribution of variant creutzfeldtjakob disease cases in the uk: what can we learn from it?*, Statistical Methods in Medical Research **12** (2003), no. 3, 235–246.

[68] Delich S and Carter AO, *Public-health surveillance - historical origins, methods and evaluation*, Bulletin of the World Health Organization **72** (1994), no. 2, 285 –304.

[69] Plog S, *Spatial organization and exchange: Archaeological survey on northern black mesa*, Southern Illinois University Press, Carbondale, 1986.

[70] Fienberg SE and Shmueli G, *Statistical issues and challenges associated with rapid detection of bio-terrorist attacks*, Statistics in Medicine **24** (2005), no. 4, 513 –529.

[71] Swan SH, Shaw G, Harris JA, and Neutra RR, *Congenital cardiac anomalies in relation to water contamination, santa-clara county, california, 1981-1983*, American Journal of Epidemiology **129** (1989), no. 5, 885–893.

[72] Mason SJ and Graham NE, *Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and in-*

terpretation, Quarterly Journal of the Royal Meteorological Society **128** (2002), no. 584, 2145–2166.

[73] Teutsch SM and Churchill RE, *Principles and practice of public health surveillance*, Oxford University Press, New York, 1994.

[74] Wood SN, *Thin plate regression splines*, Journal of the Royal Statistical Society. Series B-Methodological **65** (2003), no. 1, 95–114.

[75] Webster T, *Spatial analysis of lung, colorectal, and breast cancer on cape cod: an application of generalized additive models to case-control data*, Environmental Health: A Global Access Science Source **4** (2005), no. 11.

[76] Hastie TJ and Tibshirani RJ, *Generalized additive models*, Statistical Science **1**, no. 3.

[77] ———, *Generalized additive models*, Chapman & Hall, London, 1990.

[78] Perneger TV, *What's wrong with bonferroni adjustments*, British Medical Journal **316** (1998), no. 7139, 1236–1238.

[79] Ceccato V and Haining R., *Crime in border regions: The scandinavian case of resund, 1998-2001*, Annals of the Association of American Geographers **94** (2004), 807–826.

[80] Hardle W, Huet S, Mamen E, and Sperlich S, *Bootstrap inference in semiparametric generallized additive models*, Econometric Theory **20** (2004), 265–300.

[81] Zhou GY Ji S Wang ZL Wu J, Yuan SF and Wang Y, *Design and evaluation of a wireless sensor network based aircraft strength testing system*, SENSORS **9** (2009), no. 6, 4195–4210.

[82] Miyashita Y Yoshida M, Naya Y, *Anatomical organization of forward fiber projections from area te to perirhinal neurons representing visual long-term memory*

in monkeys, Proceedings of the National Academy of Sciences of the United States of America **100** (2003), 4257–4262.

[83] Chair Z and Varshney PK, *Optimal data fusion in multiple sensor dection systems*, IEEE Trans. Aerospace and Electronic Systems **AES-22** (1986), no. 1, 98–101.