Prospective Spatio-Temporal Surveillance Methods for the Detection of Disease Clusters

J. Brooke Marshall

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Dr. Dan J. Spitzner, Co-Chair
Dr. William H. Woodall, Co-Chair
Dr. Jeffrey B. Birch
Dr. Marion R. Reynolds

June 10, 2009
Blacksburg, VA

Keywords: ARL performance, disease surveillance, control charts, CUSUM, local Knox statistic, Poisson regression, prospective monitoring, space-time clusters, wavelets

Prospective Spatio-Temporal Surveillance Methods for the Detection of Disease Clusters

J. Brooke Marshall

ABSTRACT

In epidemiology it is often useful to monitor disease occurrences prospectively to determine the location and time when clusters of disease are forming. This aids in the prevention of illness and injury of the public and is the reason spatio-temporal disease surveillance methods are implemented. Care must be taken in the design and implementation of these types of surveillance methods so that the methods provide accurate information on the development of clusters. Here two spatio-temporal methods for prospective disease surveillance are considered. These include the local Knox monitoring method and a new wavelet-based prospective monitoring method.

The local Knox surveillance method uses a cumulative sum (CUSUM) control chart for monitoring the local Knox statistic, which tests for space-time clustering each time there is an incoming observation. The detection of clusters of events occurring close together both temporally and spatially is important in finding outbreaks of disease within a specified geographic region. The local Knox surveillance method is based on the Knox statistic, which is often used in epidemiology to test for space-time clustering retrospectively. In this method, a local Knox statistic is developed for use with the CUSUM chart for prospective monitoring so that epidemics can be detected more quickly. The design of the CUSUM chart used in this method is considered by determining the in-control average run length (ARL) performance for different space and time closeness thresholds as well as for different control limit values. The effect of nonuniform population density and region shape on the in-control ARL is explained and some issues that should be considered when implementing this method are also discussed.

In the wavelet-based prospective monitoring method, a surface of incidence counts is modeled over time in the geographical region of interest. This surface is modeled using

Poisson regression where the regressors are wavelet functions from the Haar wavelet basis. The surface is estimated each time new incidence data is obtained using both past and current observations, weighing current observations more heavily. The flexibility of this method allows for the detection of changes in the incidence surface, increases in the overall mean incidence count, and clusters of disease occurrences within individual areas of the region, through the use of control charts. This method is also able to incorporate information on population size and other covariates as they change in the geographical region over time. The control charts developed for use in this method are evaluated based on their in-control and out-of-control ARL performance and recommendations on the most appropriate control chart to use for different monitoring scenarios is provided.

Acknowledgements

A body of research of this magnitude requires patience, perseverance, and hard work to complete. It also could not be accomplished without the guidance, support, and contributions of others. I would like to thank my advisors, Dr. Bill Woodall and Dr. Dan Spitzner, for their unwavering support throughout my Ph.D. process. They taught me valuable research skills that I will take with me throughout my career, but more importantly showed me that all obstacles can be overcome with persistence. The time they spent guiding me through the research process will not be forgotten and I will always be indebted to them both. I would also like to thank Dr. Jeff Birch and Dr. Marion Reynolds for their time spent serving on my committee. Dr. Birch and Dr. Reynolds also instructed me in my undergraduate and graduate statistics programs. I appreciate all of the time and effort they put into providing me with a great statistical education, without which this research would not have been possible.

I would like to thank my family for their wholehearted support while I completed my Ph.D. In particular, I would like to thank my parents, Jan and Philip Marshall, and my grandparents, Lila and Jim Marshall. They have not only supported me as I worked toward completing my Ph.D., but have supported me through all of my academic endeavors. Both my parents and grandparents taught me the importance of education at a young age, and encouraged me to do well throughout my grade school, high school, and college careers. They also provided me with great advice at the times when I needed it most.

There are several other individuals who have made it possible for me complete my Ph.D. in Statistics, and who I would like to thank. The first is Mike Box. I appreciate all of the help he provided in the Unix computing lab in the Statistics Department at Virginia Tech. He provided me with all of the information I needed to use the

Unix computers and always made sure that I had resources available for running the simulations required for my dissertation. In addition, I would like to thank Janice McBee and Roxanne Gile for their support during my assistantship with the Virginia Tech Office of Institutional Research and Effectiveness. I would also like to thank David Radley and Lisa Lupinacci at Merck Research Labs. They were both very encouraging through the final stages of my Ph.D. process and made sure I had the time I needed to complete my dissertation, even though I had competing priorities at work.

Last but not least, I would like to thank Dr. Dan Yates, who was my high school AP Statistics instructor. His enthusiasm for statistics was engaging and I was immediately drawn to the subject of statistics because of his ability to show how useful and practical statistics can be. Had I not been introduced to this subject by such a devoted instructor, I might never have chosen statistics as my major when I enrolled as an undergraduate at Virginia Tech.

– Brooke Marshall

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In epidemiology it is important to monitor disease occurrences and their symptoms to prevent the injury and mortality of the public. Surveillance systems are implemented for this purpose and, as stated by Thacker *et al.* (1995) and Heymann and Rodier (1998), can be used to detect disease clusters and their locations, help quantify the magnitude of a disease outbreak, control or prevent the spread of disease or infection, and determine where health programs should be implemented or modified. It is the responsibility of epidemiologists and statisticians to develop these surveillance systems, and to evaluate and improve them to ensure that they are capable of producing the information needed for their intended purpose. That is the overall goal of this research.

1.1 Types of Disease Surveillance Methods

There are several types of methods used for disease surveillance which include syndromic, temporal, spatial, and spatio-temporal surveillance methods. Before an attempt can be made to develop, improve, or evaluate these types of surveillance methods, the aspects of these methods must be considered. Therefore, the main goals, advantages, and limitations of these methods are discussed. The present research focuses on spatio-temporal methods, but understanding the aspects of all of these types of methods is important for comparison.

Syndromic surveillance is used when there is a need to monitor patient complaints or symptoms of a disease for early outbreak detection. In this case, the goal is to

predict an increase in occurrences before the number of disease occurrences reaches a level where it could be detected using the more traditional temporal and spatial surveillance methods. Methods of this type are outlined in Lawson and Kleinman (2005). These methods are useful because they allow for the quickest detection of a possible outbreak. Their implementation, however, is difficult because they require the use of various types of data, including information on hospital emergency room visits and over-the-counter drug sales of remedies for particular symptoms. Shmueli and Burkom (2009) discuss some of the challenges in monitoring syndromic surveillance data.

Temporal surveillance methods are used for monitoring disease incidence counts, or other disease incidence data, for a single reporting unit or area, such as a hospital or city. These methods are discussed in Farrington and Beale (1998), Sonesson and Bock (2003), Lawson and Kleinman (2005), and Woodall *et al.* (2008). Temporal surveillance methods monitor a specified incidence statistic, which is based on the incidence data, over time and signal once a predetermined cutoff value is exceeded. A signal usually indicates an increase in the disease incidence rate, but it can also indicate a decreased rate if the method is capable of detecting rate drops. Some of these methods incorporate ways to model the dependence structure of observations over time, which is useful if there are seasonal effects present or if the disease incidence data are temporally autocorrelated. Temporal methods are sometimes used to monitor multiple reporting units, but this is not recommended as these methods do not take into account the spatial location or proximity of the reporting units, which can influence the incidence statistic being monitored.

Spatial surveillance methods have been developed to meet the goal of assessing cluster formation at some point in the past for multiple reporting units and larger geographical areas. Methods for spatial surveillance are covered in Marshall (1991) and Lawson (2006). A common goal of these methods is to detect clusters of disease occurrences within the geographical region of interest. These methods take into account the location and proximity of disease occurrences when computing the monitored incidence statistic and some methods also account for spatial autocorrelation between closely neighboring regions. Once the incidence statistics have been computed over the entire region, they are commonly displayed on a geographical map using different

colors for intervals of values of these statistics. This way the locations of extreme values can be seen easily and possible clusters of disease can be identified. In cases where it is important to determine how disease clusters form and change over time, these methods are not optimal since they do not incorporate information on the time of disease occurrences. This is usually overcome by splitting the data arbitrarily into separate time intervals and doing a separate analysis for each time period. This is not a good approach since it does not allow for detecting disease clusters at the time of occurrence or for modeling the dependence in outcomes over time.

In order to incorporate both temporal and spatial information into a surveillance system, spatio-temporal disease monitoring methods have been developed. These methods are used to find clusters of disease in both space and time, and to detect clusters in space as they are forming in time. Methods of this type have been reviewed by Williams (1984), Marshall (1991), Farrington and Beale (1998), Sonesson and Bock (2003), Lawson and Kleinman (2005), and Woodall *et al.* (2008). Spatio-temporal methods are advantageous because they combine aspects of temporal and spatial methods, and in some cases, can account for spatial autocorrelation, temporal dependence, or both. While the available methods for disease surveillance are vast, there are fewer spatio-temporal methods compared to the number of methods for monitoring diseases using spatial or temporal information alone. This is understandable given the complication of incorporating both the spatial and temporal structure of disease occurrences into a model or method; but, it makes research in the area of spatio-temporal surveillance more crucial, so that effective surveillance methods will be available for applications. Therefore, the main focus of this research is on spatio-temporal disease surveillance methods. The specific elements of spatio-temporal methods are now discussed in Section 1.2.

1.2 Elements of Spatio-Temporal Disease Surveillance Methods

When developing a spatio-temporal surveillance system, or choosing a system for a given application, several elements must be considered. Some of these elements relate

to the administration of the system, such as how and when data will be collected, while others relate to the statistical methodology implemented. Since the focus here will be on the statistical component of surveillance systems, only issues that pertain to statistical methodology will be discussed. Some of these issues are selecting the type of analysis, the level of aggregation of the available data, and the type of disease under surveillance.

The goal of the disease monitoring application will determine whether the surveillance method should be retrospective or prospective. Retrospective methods are used to determine if a disease cluster formed at some point in the past, whereas prospective methods are used to determine if a disease cluster is currently forming. In a retrospective analysis, data on disease occurrences are analyzed once at the end of a study period. In a prospective analysis, data are analyzed periodically over time as new observations are obtained. This is usually done at equal time increments. Prospective analyses are typically favored because clusters of disease can be identified as they are forming. This allows for quicker outbreak detection and for a quicker response to control the outbreak. In some cases, however, the data source will determine the type of analysis used. For instance, if disease occurrences are recorded by a third party and this information is only made available once all of the data have been collected, retrospective methods should be used. Prospective methods are most appropriate when data are available for analysis in real-time.

The type of data available will also influence the surveillance method. The data for disease surveillance have a space component, which gives information on the location of disease occurrences, and a time component, which provides information on when the occurrences took place. For each of these components the data collected may be aggregated or may have no aggregation. If the data are aggregated in space this means that one only knows an incidence occurred within a certain subregion of the region of interest and does not know the exact geographical coordinates of the occurrence. For example, if the region under surveillance is a state, the county of the occurrence may be known but the exact location may not. If the data are aggregated in time then one knows an incidence occurred within a certain window of time and does not know the exact time. When there is no aggregation in space and time then the data are referred to as point-pattern data. This type of data is becoming less common due to privacy

issues. It is more likely to see data aggregated according to political boundaries, such as census tracts or state and county borders.

Another issue that impacts the surveillance method is the type of disease being monitored. Some diseases are chronic, while others are infectious. Infectious diseases, such as influenza and malaria, are those that are capable of being spread from person-to-person. Chronic diseases are those that are ongoing or recurring and are not spread by direct person to person contact. They can be caused by heredity, lifestyle choices, or environmental factors. Chronic diseases include leukemia and heart disease. If the disease being monitored is infectious, spatial autocorrelation must be modeled to account for the spread of disease in adjacent areas of the region, and temporal autocorrelation must also be modeled in some cases to account for the spread of disease over time. Chronic diseases do not spread like infectious diseases since they are noncommunicable. Therefore, in cases of chronic disease surveillance, each individual has an independent chance of acquiring the disease and it is not necessary to model spatial or temporal autocorrelation. With chronic diseases, however, the influence of environmental factors and other covariates, such as age and gender, should be taken into account. Since these covariates can affect the probability of obtaining the disease, a method that can incorporate different baseline rates of disease over the region are more useful. Time dependency due to seasonal or other trends can be present for cases of both chronic and infectious diseases. When these trends are present, time series methods should be used so that these trends can be modeled.

1.3 Overview of Topics

The topics covered in the subsequent chapters relate to the development, evaluation, and improvement of spatio-temporal disease surveillance methods. There will be an emphasis on prospective spatio-temporal methods, as opposed to retrospective methods, because these methods are more advantageous due to their ability to detect clusters of disease more quickly. Retrospective methods will only be mentioned in relation to the prospective methods discussed. The prospective spatio-temporal methods considered are those used for monitoring chronic diseases using both aggregated and point-pattern data, taking covariates into consideration.

In Chapter 2, an evaluation of the local Knox monitoring method of Rogerson (2001) is given, which was performed by Marshall *et al.* (2007). This method was designed for monitoring chronic diseases when point pattern data are available. The method detects clusters of disease through the use of a cumulative sum (CUSUM) chart, which is used to monitor the mean of the local Knox statistic. A detailed description of this method is given and the effects that the population density and the shape of the geographical region under surveillance have on the performance of this method are shown. The distribution of the local Knox statistic is also examined.

In Chapter 3, a wavelet-based prospective surveillance method for monitoring chronic diseases is developed. This method can be used under the assumptions that aggregated incidence count data are available and that the incidences within each subregion follow an independent Poisson process over time. The wavelet-based method uses Poisson regression, with wavelets as regressors, to model the spatial aspects of disease incidences. One of three possible control charts is then used to monitor parameters of this model and to detect forming clusters of disease. If the control chart produces a signal, diagnostic tools have been developed that can be used to determine the size and location of the disease cluster or clusters that have formed.

In Chapter 4, the wavelet-based disease surveillance method developed in Chapter 3 is evaluated. The evaluation focuses on the in-control and out-of-control average run length (ARL) performance of the control charts used in the wavelet-based method. The ARL performance of these control charts is considered for many scenarios to determine how often false signals are produced when there are no disease clusters present and how quickly the charts detect clusters when they truly are present. A case study is also provided, where the method is used to monitor female respiratory lung cancer incidences in the state of New Mexico.

Finally, in Chapter 5, the evaluation of the Rogerson (2001) method, the development of the wavelet-based disease surveillance method, and the evaluation of the wavelet-based method are summarized. The advantages and disadvantages of these methods are given. In addition, areas of future research related to these methods will be discussed.

Chapter 2

Use of the Local Knox Statistic for the Prospective Monitoring of Disease Occurrences

2.1 Introduction

A common problem in the area of public health surveillance is the detection of disease outbreaks or epidemics in a geographic region. One way to address this problem is to look for the clustering of outcomes of disease in time and space. In this context, a cluster is a group of disease occurrences where events that happen close in time also have a tendency to happen close in space. This is referred to as space-time interaction.

The methods generally used to detect whether space-time interaction has occurred are those of Knox (1964), Mantel (1967), and Jacquez (1996). These methods test for space-time interaction using point-pattern data. Point pattern data consist of observations where events happen randomly in space and the exact coordinates and time of each event are observed. To determine whether events interact in space and time, spatial and temporal distances between pairs of points are examined. To calculate the Knox (1964) statistic, one compares all pairs of observations and then counts the number of pairs of occurrences close in both space and time. The k nearest neighbor test, developed by Jacquez (1996), is also based on a count. With Mantel's (1967) test, one takes the product of the space and time distances between two points and sums these products for all pairs.

The Knox (1964), Mantel (1967), and k-nearest neighbour (1996) methods are all retrospective methods used for the surveillance of space-time interaction since they compare all pairs of points once, at the end of a study period. It is typically more desirable to detect disease outbreaks as they are occurring in real-time so that preventative action can be taken. Therefore, a shift should be made from the use of retrospective surveillance methods to prospective methods. With prospective surveillance, testing is done sequentially as each new event occurs instead of only at the end of a study. This allows for quicker detection of disease clusters.

Sonesson and Bock (2003), Lawson and Kleinman (2005), and Woodall (2006) have discussed methods of prospective surveillance in public health applications. Prospective surveillance methods for detecting space-time clusters in point pattern data have been developed by both Kulldorff (2001), using a Bernoulli model, and Rogerson (2001). Kulldorff (2001) used a prospective space-time scan statistic to test for clusters as new observations are obtained. Rogerson (2001) used the cumulative sum (CUSUM) control chart to monitor a local Knox statistic, which is based on the statistic proposed by Knox (1964).

The CUSUM chart, used in the prospective monitoring of the local Knox statistic, is a tool used often in industrial statistical process control (SPC). The CUSUM chart has two parameters, a control limit and a reference value. In SPC, these parameters are chosen so that the CUSUM charts will have acceptable average run length (ARL) performance under a specified in-control distribution of the statistic being monitored. The run length in our case is the number of disease occurrences until the chart signals that a space-time interaction is present.

The design of the CUSUM chart has been studied extensively in SPC applications, but it has rarely been discussed in public health applications. The CUSUM chart has been used to detect patterns of disease occurrence in the methods developed by Raubertas (1989), Rogerson (1997), Leung *et al.* (1999), Järpe (1999), Rogerson and Yamada (2004), and Sonesson (2007) in addition to the local Knox monitoring method. Joner *et al.* (2008) studied the ARL performance for the method developed by Rogerson and Yamada (2004), but the performance of the CUSUM chart for monitoring the local Knox statistic has not previously been investigated.

It is important to understand the performance of CUSUM charts when designing them for specific public health monitoring applications, as Lawson (2001) pointed out. Here we discuss the design of the CUSUM control chart for monitoring the local Knox statistic by examining in-control ARL performance for varying values of the space and time thresholds and the CUSUM control limit. We also determine the effect of population density and region shape on ARL performance. We demonstrate that the ARL performance of this chart is highly influenced by the threshold values, population density, and region shape because these factors alter the distribution of the local Knox statistic. This makes designing the CUSUM chart impossible without computer simulation. We also discuss how the ARL performance deviates widely from the ARL performance under an assumed normal in-control distribution of the monitored statistic and examine the properties of the true in-control distribution.

2.2 Review of the Local Knox Monitoring Method

2.2.1 The Knox Test and Local Knox Test

The Knox test is frequently used to test for interactions in space and time. This test has a natural application in epidemiology when there is interest in determining whether disease occurrences have happened with a higher frequency than expected in a limited spatial area. To apply the Knox test, the time of onset of a disease must be known, as well as the coordinates of where the disease occurred, for n events. Therefore, a single observation or point consists of three values, the time the event occurred and the longitude and latitude where the event occurred.

The Knox statistic, denoted by N_{st}, is a count of the number of pairs of the n points that are close in space and time out of the $n(n-1)/2$ possible pairs. A pair of points are considered close in space if the Euclidean distance between the two events is less than a given space threshold, s. Likewise, a pair of points are considered close in time if the time between the two events is less than a given time threshold, t. Then a pair of points are close in both space and time if both of these criteria are met. The number of pairs of points close in space, time, and space and time are denoted by n_s, n_t, and n_{st}, respectively.

There are four hypothesis testing procedures for this statistic, which are used to determine whether there is significant space-time interaction. These methods have been outlined by Kulldorff and Hjalmars (1999) and Rogerson (2001). Three of these methods use different parametric null distributions to determine a p-value for the statistic and the fourth uses a Monte Carlo approach for obtaining the null distribution by permuting the times over the fixed spatial locations.

In order to determine which observations are contributing to a significant space-time interaction, Rogerson (2001) developed a local Knox test, which works by evaluating space-time interaction near an individual point. For a specific observation i, $n_s(i)$ is the number of observations close to observation i in space based on the threshold value s. Similarly, $n_t(i)$ is the count of observations close in time to point i with respect to the threshold t and $n_{st}(i)$ is the count of observations close to point i in both space and time. The local Knox statistic for observation i is $N_{st}(i)$, which takes the value $n_{st}(i)$.

The null distribution of the local Knox statistic can be modeled by a mixture of n hypergeometric distributions. To construct this distribution, consider a pair of observations i and j. Given the spatial location of observation i and the time of observation j, the probability distribution of $N_{st}(i)$ is hypergeometric with parameters $n-1$, $n_s(i)$, and $n_t(j)$. Under the null hypothesis we assume that each time, out of the n times at which an incidence is observed, has the same probability of being observed at the location of observation i, which implies that each permutation of times over the fixed spatial locations is equally likely. There are a total of $n!$ permutations of times over the fixed spatial locations. When a pair of observations i and j are held constant, there are then $(n-1)!$ ways to permute the remaining times over the other fixed spatial locations. This implies that for each $j = 1, 2, \ldots, n$ there are $(n-1)!$ permutations of the total $n!$ permutations that give the same hypergeometric distribution for $N_{st}(i)$. So for the fixed location of observation i, the value $n_{st}(i)$ has an equal chance of coming from any one of the $n!/(n-1)! = n$ hypergeometric distributions with parameters $n-1$, $n_s(i)$, and $n_t(j)$, for $j = 1, 2, \ldots, n$. Therefore, the distribution of $N_{st}(i)$, for a fixed spatial location, can be modeled by a mixture of the n hypergeometric distributions with probability mass function

$$P\left(N_{st}(i) = n_{st}(i)\right) = \frac{1}{n}\sum_{j=1}^{n} \frac{\binom{n_t(j)}{n_{st}(i)}\binom{n-1-n_t(j)}{n_s(i)-n_{st}(i)}}{\binom{n-1}{n_s(i)}}. \quad (2.1)$$

The expectation of the local Knox statistic and the exact variance of the statistic are

$$E[N_{st}(i)] = \frac{2n_t n_s(i)}{n(n-1)} \quad (2.2)$$

and

$$V[N_{st}(i)] = \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\left(n-1-n_s(i)\right)}{n(n-1)^2(n-2)} + \frac{n_s(i)^2 \sum_{j=1}^{n}\left[n_t(j) - \frac{2n_t}{n}\right]^2}{n(n-1)^2}, \quad (2.3)$$

respectively. The value n_t is equal to $\frac{1}{2}\sum_{j=1}^{n} n_t(j)$, which is also the number of pairs of points close in time for the Knox test. The derivation of the variance in equation (2.3) is shown in Appendix A. Rogerson (2001) lists the same expected value, but states the variance of $N_{st}(i)$ as

$$V[N_{st}(i)] \cong \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\left(n-1-n_s(i)\right)}{n(n-1)^2(n-2)}. \quad (2.4)$$

Notice that the first term in the exact variance formula in equation (2.3) is identical to equation (2.4); hence, the variance in equation (2.4) can be viewed as an approximation of the true variance. The second term in equation (2.3) will always be nonnegative. Consequently, the approximated variance will be less than or equal to the true variance in all cases.

To determine the approximate significance of the local Knox statistic, Rogerson (2001) suggested approximating the distribution of $N_{st}(i)$ with a normal distribution. The expectation and variance used for this normal approximation are the expectation and variance given in equations (2.2) and (2.4), respectively. Standardizing $N_{st}(i)$ gives the statistic for observation i, which is

$$z_{st}(i) = \frac{n_{st}(i) - E[N_{st}(i)] - 0.5}{\sqrt{V[N_{st}(i)]}}, \tag{2.5}$$

where 0.5 is included as a continuity correction. Rogerson (2001) assumed that this standardized local Knox statistic has an approximate standard normal distribution. An alternative approximation could also be used for this test where the approximate variance is replaced by the exact variance given in equation (2.3).

2.2.2 Prospective Monitoring with the Local Knox Statistic

The local Knox statistic can be used to monitor disease occurrences prospectively with a CUSUM chart. The CUSUM control chart is a tool that is often used in SPC to monitor industrial processes, as has been discussed by Hawkins and Olwell (1998). The most common version of this type of control chart is designed to detect changes in the mean of a random variable that has a normal distribution with an in-control mean μ_0 and standard deviation σ. This chart is particularly useful, compared to other control charts, when the random variable being monitored is a single observation. In our application, this variable is $z_{st}(i)$.

The idea behind the CUSUM chart is to monitor the sum of the deviations of each observation from the in-control mean μ_0, which can be used to detect an increase or decrease in the mean of the observations. When there is interest in detecting an

increase in the mean, the value plotted on the chart for each observation is

$$C_i = max\left(0, x_i - \mu_0 - K + C_{i-1}\right), \tag{2.6}$$

where $C_0 = 0$. In equation (2.6), x_i is the value of the statistic being monitored, K is a reference value, and C_{i-1} is the CUSUM value for observation $i-1$. The value K is one-half of the smallest increase in the target mean one wishes to detect quickly and is typically defined as $K = k\sigma$. In most cases $k = \frac{1}{2}$ is used for these charts because it has been shown that this value will lead to good statistical performance.

To determine when the mean has shifted, the value C_i is compared to $H = h\sigma$. If $C_i \geq H$ then the chart signals, indicating that the mean has increased. The value of H is chosen based on desired in-control ARL performance.

Rogerson (2001) proposed the use of the CUSUM chart to monitor changes in the Knox statistic each time a new event occurs. A signal for this chart would indicate that we have moved from an in-control case of no space-time interaction in disease occurrences, to the existence of space-time interaction, which corresponds to the out-of-control case. To evaluate a change in the Knox statistic from observation $i-1$ to observation i, the Knox statistic calculated for observation i, which is denoted by $N_{st}(1:i)$, is compared to the conditional distribution of $N_{st}(1:i)$ given $N_{st}(1:i-1)$, $n_s(i)$, and $n_t(j)$ for $j = 1, 2, \ldots, n$. The statistics used in this chart are

$$z_i = \frac{N_{st}(1:i) - E[N_{st}(1:i)|N_{st}(1:i-1), n_s(i), \{n_t(j)|j = 1, \ldots, n\}] - 0.5}{\sqrt{V[N_{st}(1:i)|N_{st}(1:i-1), n_s(i), \{n_t(j)|j = 1, \ldots, n\}]}} \tag{2.7}$$

for $i = 1, 2, \ldots, n$, which are assumed in Rogerson (2001) to follow an approximate standard normal distribution. There, it is also shown that this statistic is equivalent to the local Knox statistic defined in equation (2.5). This result holds for use of either the variance in equation (2.3) or that in equation (2.4). Therefore, the standardized local Knox statistic, $z_{st}(i)$, can be used to monitor changes in the Knox statistic over

time. This is done by replacing n in equations (2.2) and (2.4) with the most recent observation number. Since the standardized local Knox statistic is assumed to have an approximate standard normal distribution and $k = \frac{1}{2}$, the CUSUM values plotted for this chart are

$$C_i = max\left(0, z_{st}(i) - \frac{1}{2} + C_{i-1}\right), \tag{2.8}$$

for $i = 1, 2, \ldots, n$, where $C_0 = 0$. This chart will signal, indicating the existence of space-time interaction, when $C_i \geq h$.

2.2.3 Application of the Local Knox Monitoring Method

Rogerson (2001) provided an application of the prospective local Knox monitoring method using data collected by Williams *et al.* (1978) on Burkitt's lymphoma incidences in the West Nile district of Uganda from 1961–1975. These data had been previously analyzed by Williams *et al.* (1978) using the Knox test. A total of 202 cases were recorded over this time period, and the data obtained on these cases included the geographical coordinates of the home of each patient and the date of onset of the disease for each patient. Of these 202 observations, only 188 had information on both the location of the patient's home and the date of disease onset. The geographical coordinates of these 188 cases are shown in Figure 2.1.

In order to detect space-time clusters of disease within this district of Uganda, Williams *et al.* (1978) split the 188 cases into three subsets for the five year increments 1961–1965, 1966–1970, and 1971–1975. For each of these subsets, the significance of the Knox test was determined for 30 space and time threshold combinations. Each of these combinations had a space threshold value of $s = 2.5$, 5, 10, 20, or 40 km and time threshold value of $t = 30$, 60, 90, 120, 180, or 360 days. For the period 1961–1965, the Knox statistic was significant for 19 threshold combinations. These included $s = 2.5$ and $t = 180$, $s = 5$ and $t = 180$ and 360, $s = 10$ and all values of t, $s = 20$ and $t = 30$, 60, 180, and 360, and $s = 40$ and all values of t. There were only two significant results for time period 1966–1970, which were for threshold combinations $t = 360$ and $s = 10$ and

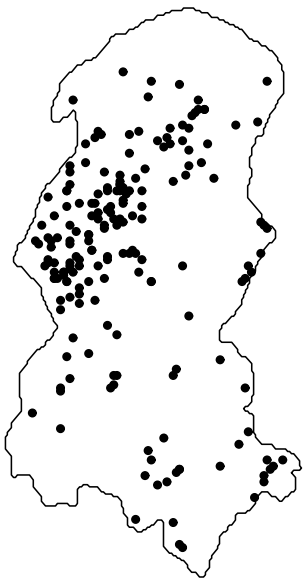


Figure 2.1: Burkitt's lymphoma incidences in the West Nile district of Uganda from 1961–1975

20. From 1971–1975, the test statistics that indicated significant space-time clustering were for threshold combinations $s = 10$ and $t = 90$, 120, and 180. Based on these results, Williams *et al.* (1978) concluded that there was strong evidence of clustering from 1961–1965. A second analysis was done using the data from 1966–1975, by further subdividing the observations collected over this time period. The data collected from 1966–1975 was subdivided into five subsets, which each covered two years. This second analysis showed strong evidence of space-time clustering from 1972–1973.

Rogerson (2001) used the same 188 cases of Burkitt's lymphoma to illustrate the local Knox monitoring method. Rogerson (2001) analyzed these data prospectively

as described in Section 2.2.2 for each of the space and time threshold combinations used by Williams *et al.* (1978). The control limit used for the CUSUM charts in this application was $h = 5.64$, which was determined using the approximation of Siegmund (1985) to achieve a 0.10 false alarm probability. The CUSUM charts signaled two or more times, indicating significant space-time clustering, for space and time threshold combinations $s = 2.5$ and $t = 90$, $s = 10$ and $t = 180$ and 360, $s = 20$ and $t = 90$, 120, 180, and 360, and $s = 40$ and $t = 60$, 90, 120, 180, and 360. The control charts for these threshold combinations signaled at different time points, indicating clustering at different times over the period 1961–1975.

The Burkitt's lymphoma results of Rogerson (2001) are inaccurate due to the fact that the approximate variance in equation (2.4) was used instead of the exact variance in equation (2.3), and because errors were discovered in the computation of the CUSUM values plotted on the control charts. Therefore, the analysis was redone to give the corrected results of this method when applied to the Burkitt's lymphoma disease incidence data. The analysis of Rogerson (2001) was redone using the same space and time threshold combinations and the same control limit, $h = 5.64$, for the CUSUM charts. The CUSUM values were computed using the exact variance in equation (2.3). The corrected CUSUM charts for each space and time threshold are shown in Appendix B. None of the CUSUM charts produce a signal for any of the space and time threshold combinations. These new results indicate that there was either no space-time clustering of Burkitt's lymphoma over the time period 1961–1975, or that the CUSUM chart was designed incorrectly. The results in Section 2.3 show that these results are due to poor design.

2.3 CUSUM Design for the Local Knox Monitoring Method

2.3.1 ARL Performance

When control charts are used for industrial process monitoring, the charts are often designed based on ARL performance. When designing a CUSUM chart, the in-control ARL must be considered as a function of the control limit H so that the appropriate

value of H is used to achieve a specified ARL. The in-control ARL of a CUSUM chart is the ARL when there has been no shift in the mean of the monitored statistic from the target value. In general, when the process is in-control, large ARLs are preferable since a signal in this case would be a false alarm. If, however, the chart is designed to achieve a very large in-control ARL, then it will also be harder for the chart to detect true changes in the mean of the monitored statistic from a target value. The in-control ARL performance of the CUSUM chart used for monitoring the standardized local Knox statistic is considered in this sub-section for two values of h, 2.5 and 3.0, under various specifications of s and t. The ARL performance of this chart is investigated for an in-control case where no space-time interaction exists and the population density is uniform within a region.

To simulate the in-control state, a unit square region was defined where longitude defines the horizontal axis and latitude defines the vertical axis. The spatial location of an observation was obtained by randomly generating coordinates uniformly on the unit square. Times between disease occurrences were modeled by an exponential distribution with mean equal to 1. The mean value can represent any unit of time depending on the application. After each observation was obtained, the local Knox statistic and CUSUM values were calculated. The CUSUM value, C_i, was then compared to the control limit, h. This process was repeated until $C_i \geq h$ and the chart signaled, producing one run length. To estimate the ARL, 80,000 run lengths were obtained under these same conditions and then averaged.

To study how the thresholds s and t affect the ARL, different combinations of these thresholds were selected and the ARL was estimated for each case under the scheme above. Four time thresholds and four space thresholds were chosen giving a total of 16 threshold combinations considered. The time thresholds chosen were 0.5, 1, 2, and 6. These values of t are -0.5, 0, 1, and 5 standard deviation units from the mean time between occurrences, respectively. The space threshold is a measure of Euclidean distance. Accordingly, the space thresholds were selected so that a circle with radius s would cover 5%, 10%, 25%, or 50% of the area in the unit square region if the circle was completely contained within the region. The values of s that meet these criteria are 0.1262, 0.1784, 0.2821, and 0.3989, respectively.

Another factor to consider in the ARL performance of this chart is the difference in in-control ARLs when the exact variance in equation (2.3) or approximate variance in equation (2.4) is used in the calculation of $z_{st}(i)$, which is then used in the calculation of the CUSUM values. To compare the in-control ARL performance of these variances, the ARLs were computed using the two different variance expressions for each of the 16 space and time threshold combinations.

Table 2.1 contains the in-control ARL simulation results. The standard errors, SE_{ARL}, are also given to show the precision of the ARL estimates. These results show that the in-control ARL performance changes depending on the variance formula used. The ARLs are consistently larger when using the exact variance in equation (2.3) because it is always greater than or equal to the approximate variance in equation (2.4). There is no obvious pattern indicating how changes in the control limit and the space and time threshold values impact in-control ARL performance. Notice that for a specific value of h, the ARLs change dramatically for different threshold combinations and the pattern of ARLs across the different space thresholds is different for each time threshold. This is illustrated in the plots of ARL values in Figure 2.2. In the case when $h = 2.5$ and the exact variance expression (2.3) is used, Figure 2.2 (a) shows that when $t = 0.5$, the ARLs reach a peak and then begin to decrease as the space threshold increases; but, when $t = 6$ the ARLs increase monotonically as the space threshold increases. When the value of h increases, the ARLs increase as expected, but the relationship between ARLs for different values of s and t changes for the different values of h.

The in-control ARL performance varies for different threshold combinations because the distribution of $N_{st}(i)$ changes when the space and time thresholds change. This in turn changes the distribution of the monitored statistic, $z_{st}(i)$, for different space and time thresholds. Since the in-control ARL performance changes for different values of these parameters and the relationship between the ARLs for these thresholds change for each value of h, it is difficult to determine a value of h to achieve a specified in-control ARL in practice. This limits the applicability of this method of surveillance because the control chart is difficult to design. To design this CUSUM chart, the in-control ARL performance would need to be determined for multiple values of the control limit for each individual space and time threshold combination through computer simulation.

Table 2.1: Estimated in-control ARL performance results for different values of the control limit h, space and time thresholds s and t, and variance expressions in equations (2.3) and (2.4) within the uniform population density unit square region

			$t = 0.5$		$t = 1$		$t = 2$		$t = 6$	
VAR	h	s	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}
Eq.	2.5	0.1262	723.6	3.05	893.2	3.27	592.7	1.76	1269.1	4.17
(2.3)		0.1784	935.0	3.47	527.6	1.65	849.1	2.84	1746.4	5.99
		0.2821	585.7	2.02	803.3	2.80	1026.2	3.60	2962.1	10.54
		0.3989	754.7	2.60	823.9	2.83	1219.7	4.29	5170.4	18.55
	3.0	0.1262	1494.1	5.79	1042.7	3.56	1265.8	4.57	2309.7	7.76
		0.1784	1082.4	3.80	1055.6	3.75	1573.7	5.62	3190.1	10.96
		0.2821	1063.6	3.72	1278.4	4.39	1669.9	5.85	5181.0	18.21
		0.3989	1225.8	4.25	1343.9	4.64	1979.0	6.93	8681.5	30.89
Eq.	2.5	0.1262	673.0	2.84	875.9	3.23	538.8	1.57	1162.0	3.81
(2.4)		0.1784	882.9	3.30	439.6	1.32	738.8	2.47	1430.5	4.83
		0.2821	361.0	1.16	546.8	1.87	676.7	2.31	1900.8	6.75
		0.3989	403.2	1.47	389.4	1.33	574.7	2.02	2266.6	8.18
	3.0	0.1262	1367.6	5.20	1013.5	3.48	1106.9	3.91	2077.7	7.01
		0.1784	1057.1	3.78	819.9	2.83	1257.8	4.31	2630.9	9.11
		0.2821	659.4	2.29	897.5	3.15	1123.0	3.94	3200.8	11.34
		0.3989	588.8	2.05	588.5	2.02	865.3	3.05	3569.2	12.82

Furthermore, if ARL performance is determined by simulation, the exact variance in equation (2.3) should be used over the approximate variance in equation (2.4) since the results using these two expressions differ substantially in many cases.

2.3.2 Effect of Population Density

The space and time thresholds used to calculate the local Knox statistic and the control limit of the CUSUM chart are not the only factors that have an impact on ARL performance. The underlying population density in a region also influences the in-control ARL values. We now discuss the effect of population density on the local Knox statistic and use simulation to show how this changes in-control ARL performance.

The effect of the population distribution is an issue often discussed in relation to the Knox statistic. According to Knox (1964), the validity of the Knox statistic is not influenced by the population density itself or by changes in the size of the population of a region over time. It is only affected by changes in the population density over

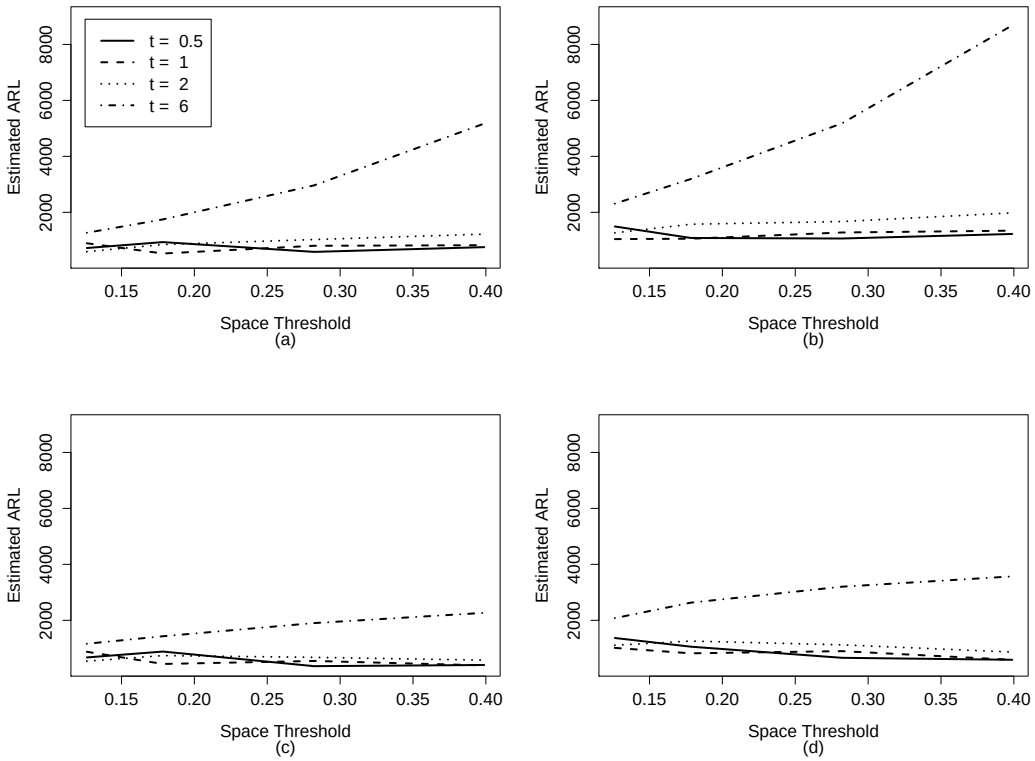


Figure 2.2: Estimated in-control ARL performance of the CUSUM chart for the 16 space and time threshold combinations when (a) $h = 2.5$ and the exact variance in equation (2.3) is used; (b) $h = 3.0$ and the exact variance in equation (2.3) is used; (c) $h = 2.5$ and the approximate variance in equation (2.4) is used; (d) $h = 3.0$ and the approximate variance in equation (2.4) is used

the study period. As long as the population distribution remains constant during the period of observation, the distribution of distances between disease occurrences will also remain constant. If this distribution is stable then any changes in the distribution can be attributed to a possible space-time interaction. If, however, the population distribution changes within different areas of a region over time, then there will be confounding between space-time interaction and nonuniformity of population density. It would not be possible to tell whether a significant test statistic is due to the existence of space-time interaction or a change in the population density.

The local Knox statistic also has a problem with confounding when the population distribution changes over time. The major assumption made when constructing the null distribution for this statistic is that each time has the same probability of being observed at a specific spatial location. If the population density changes over time, then for a fixed location the probability of times occurring there are not equally likely. Therefore, if the population density changes over time, space-time interaction will be confounded with changes in the population density for the local Knox statistic. If covariate information on changes in the population density over time is available for the monitoring period, however, the local Knox statistic could possibly be modified to account for these changes by changing the weights of the coefficients in the density function.

When monitoring the standardized local Knox statistic, not only are there concerns with changes in the population density over time, but also with the population density function itself. With the Knox statistic one considers all pairs of points at once, but with the local Knox statistic one only considers pairs of points associated with one point at a time. When considering all pairs at once, there is only one distribution of distances between observations. When considering each point individually, the distribution of distances between points changes for each point observed. This implies that there are n distributions of the local Knox statistic to consider. These distributions can vary widely depending on the population density. If a point is located in a heavily populated area of the region, then the distribution of distances between this point and all the others will be right-skewed since there will be more observations close to this point. On the other hand, if a point is located in a sparsely populated area, the distribution of distances from this point will be skewed left because there are only a few observations close to this point and more points further away in highly populated areas. Because the distance distributions of the points observed change for different population densities, the local Knox statistic is influenced by the population distribution of the region. Therefore, even if the population density remains constant over the monitoring period, the ARL performance of the CUSUM chart used to monitor the standardized local Knox statistic will change for different population densities.

To illustrate how differences in the population density can influence the in-control ARL performance of the CUSUM control chart, in-control ARLs were estimated for

a nonuniform population density to compare to the unit square uniform population density results in Section 2.3.1. The nonuniform population distribution was selected to cover the same unit square region considered previously. The unit square was split into nine subregions where the population density is uniform within each of the subregions. This distribution is represented in Figure 2.3. The value shown for each subregion is the probability of an observation falling within the subregion. The center subregion covers the area from 0.35 to 0.65 on both the horizontal and vertical axis and the probability of falling in this subregion is 0.40. The four subregions located in the corners of the region each have area 0.1225 and the probability of a point falling in one of these regions is 0.034. The subregions on the edges have area 0.105 with probability 0.116. Thus, the population density is highest in the center of the region and lowest in the corners of the region.

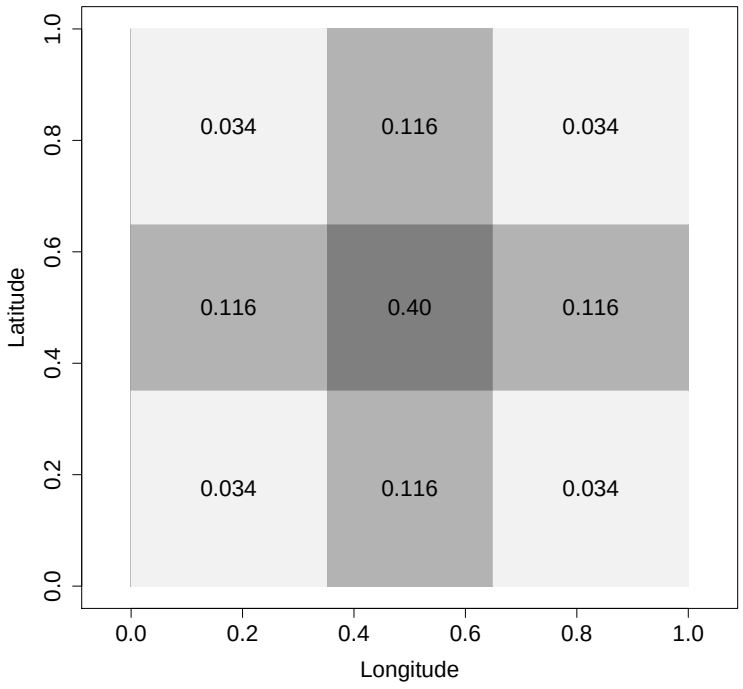


Figure 2.3: Nonuniform population distribution used for determining in-control ARL performance

The procedure for determining the in-control ARL performance for this population density was the same procedure used for the init square uniform density case. The times between occurrences were randomly generated according to an exponential distribution with mean equal to 1. All of the space and time thresholds considered in the unit square uniform density case were also considered here and the control limits selected were $h = 2.5$ and $h = 3.0$. Only the exact variance in equation (2.3) was used in the calculation of $z_{st}(i)$. The estimated in-control ARLs for the nonuniform population distribution are presented in Table 2.2.

Table 2.2: Estimated in-control ARL performance of the CUSUM chart within the nonuniform population density region for two values of the control limit h and for different space and time thresholds s and t

		$t = 0.5$		$t = 1$		$t = 2$		$t = 6$	
h	s	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}
2.5	0.1262	466.3	1.57	663.9	2.24	752.8	2.46	1503.0	4.94
	0.1784	657.7	2.30	662.9	2.23	946.3	3.29	2188.8	7.58
	0.2821	628.7	2.13	815.2	2.81	1142.6	3.95	4419.7	15.73
	0.3989	727.1	2.50	883.8	2.99	1457.2	5.07	9429.4	33.34
3.0	0.1262	768.2	2.62	1047.3	3.54	1315.4	4.53	2675.7	9.01
	0.1784	1025.9	3.53	1087.1	3.73	1609.1	5.60	3995.9	13.97
	0.2821	974.3	3.28	1382.0	4.89	1878.1	6.51	7723.6	27.21
	0.3989	1169.2	3.95	1386.8	4.71	2281.2	7.91	15262.6	54.16

The in-control ARL performance of the CUSUM chart for the uniform and nonuniform population densities is considerably different, and there does not appear to be a clear connection between the in-control ARL values for these two population densities. Plots of the in-control ARLs for each time and space threshold combination when $h = 2.5$ and $h = 3.0$ are displayed in Figure 2.4. Plot (a) of Figure 2.4 shows the difference in in-control ARL performance of the two population densities when compared to the plot of estimated in-control ARLs in Figure 2.2 (a). The nonuniform population density has a much wider range of in-control ARL values than the uniform density over the different space and time thresholds. Also, the trends of the in-control ARL values for the time thresholds across the four space thresholds change depending on the population density. As the space threshold increases, there is an overall increase in the estimated in-control ARLs for the nonuniform population density. This is not the case for the uniform population density over the unit square region. In the unit

square uniform density case, the estimated in-control ARLs increase and decrease over the range of space thresholds for small values of t.

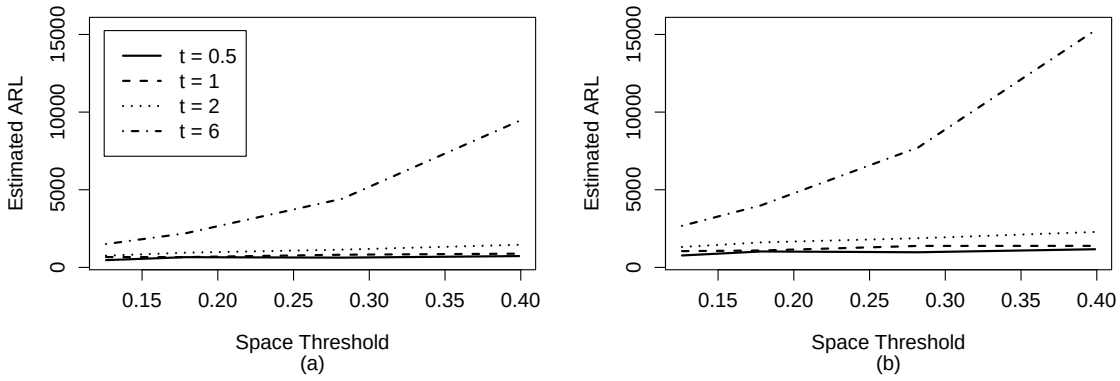


Figure 2.4: Estimated in-control ARL performance of the CUSUM chart within the nonuniform population density region when (a) $h = 2.5$; (b) $h = 3.0$

When monitoring the standardized local Knox statistic using the CUSUM chart, the in-control ARL performance will not only change when the population distribution changes over time, but also for different population distributions. Since the ARL performance changes for different population densities as well as for different space and time thresholds, it is even more difficult to determine the appropriate control limit for a specified in-control ARL. To design this CUSUM chart, information on the population density would have to be incorporated when randomly generating observations in a simulation to determine the in-control ARL values.

2.3.3 Effect of the Region Shape

The effect of the shape of the region on the in-control ARL performance of the CUSUM chart monitoring $z_{st}(i)$ is closely related to the effect of the population density. Since the local Knox statistic is affected by the population density, which depends on the shape of the region, it follows that the local Knox statistic will also be influenced by region shape. For differently shaped regions, the population density will change and the distribution of the distances from each point to other points will change for

each observation because the region boundaries are different. Since changes in the distance distributions of the local Knox statistics affect ARL performance, it is clear that the ARL performance for the CUSUM chart monitoring the standardized local Knox statistic will change for differently shaped regions.

To show how the shape of the region can influence ARL performance, in-control ARLs were estimated for a rectangularly shaped region with uniform population density, to compare with the in-control ARLs for the unit square region with uniform population density. The rectangular region was chosen to have area equal to 1, as does the unit square. This rectangular region covered the interval from 0 to 0.5 on the horizontal axis, representing longitude, and the interval from 0 to 2 on the vertical axis, representing latitude. The in-control ARL performance for the rectangular region was determined using the same procedure used for the unit square uniform density case and the nonuniform density case. The control limits used were $h = 2.5$ and $h = 3.0$ and all of the space and time threshold combinations from the unit square uniform density case and the nonuniform density case were considered. The times between occurrences were still randomly generated from an exponential distribution with mean equal to 1 and the exact variance in equation (2.3) was used to calculate $z_{st}(i)$. The estimated in-control ARLs for the rectangular region are shown in Table 2.3.

Table 2.3: Estimated in-control ARL performance of the CUSUM chart within a rectangular region for two values of the control limit h and for different space and time thresholds s and t

		$t = 0.5$		$t = 1$		$t = 2$		$t = 6$	
h	s	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}	ARL	SE_{ARL}
2.5	0.1262	679.6	2.85	921.6	3.38	588.8	1.74	1258.8	4.08
	0.1784	974.0	3.65	516.5	1.60	868.7	2.95	1684.4	5.77
	0.2821	511.5	1.69	791.9	2.72	966.8	3.37	2700.1	9.60
	0.3989	790.9	2.78	764.8	2.68	1112.9	3.91	4074.6	14.61
3.0	0.1262	1446.2	5.85	1078.1	3.71	1168.3	4.03	2248.8	7.49
	0.1784	1165.4	4.12	946.2	3.23	1543.3	5.44	3059.2	10.56
	0.2821	972.5	3.44	1307.0	4.57	1627.2	5.71	4711.7	16.63
	0.3989	1208.0	4.13	1270.0	4.43	1782.7	6.24	6832.0	24.29

The in-control ARL performance for the rectangular region differs from the in-control ARL performance of the unit square uniform density region. Plots of the in-control ARL values for the rectangular region are given in Figure 2.5. These plots

show that the trends in in-control ARL performance for each time threshold over the space thresholds for the rectangular region are similar to the trends in Figure 2.2 for the unit square uniform population density region. Even though the in-control ARL trends are similar for the rectangular and unit square uniform density cases, the ARL values still differ for many threshold combinations. Therefore, the shape of the region must be considered in the control chart design. This would be done automatically in a simulation incorporating population density to determine in-control ARL values as discussed in Section 2.3.2.

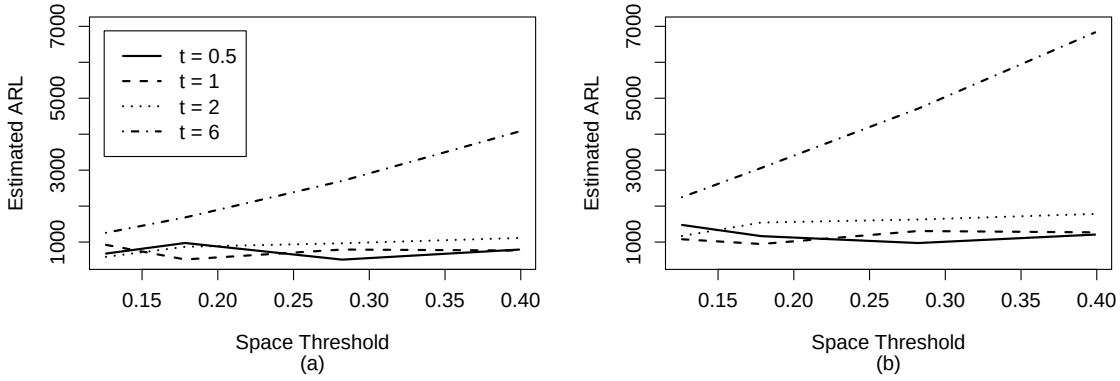


Figure 2.5: Estimated in-control ARL performance if the CUSUM chart within a rectangular region when (a) $h = 2.5$; (b) $h = 3.0$

2.3.4 Assessment of the Normal Approximation

The CUSUM chart used to monitor the standardized local Knox statistic, $z_{st}(i)$, is commonly used to monitor statistics or observations that follow a normal distribution. Rogerson (2001) stated that the distribution of the standardized local Knox statistic is non-normal and skewed right, but chose the control limit h based on a normal approximation. To directly examine the effect of the normality assumption on the design of the CUSUM chart for monitoring $z_{st}(i)$, the estimated in-control ARL values obtained in Sections 2.3.1 and 2.3.2 are compared to the in-control ARLs of the CUSUM

chart assuming one is monitoring normally distributed statistics. We also compare the properties of the distribution of $z_{st}(i)$ to the standard normal distribution.

The in-control ARL values for a one-sided CUSUM chart used when monitoring normally distributed statistics are 68.2 and 117.6 for $h = 2.5$ and $h = 3.0$, respectively. These in-control ARLs, recommended by Rogerson (2001), were computed using the method of Vance (1986). For each value of the control limit, the in-control ARLs expected under normality are much lower than the simulated ARL values in Tables 2.1 and 2.2, for all space and time threshold combinations. The control limit for a CUSUM chart is selected based on a specified in-control ARL under an assumed in-control distribution. Therefore, if the control limit for monitoring $z_{st}(i)$ values is selected under the assumption of a standard normal distribution, the true in-control ARL of the chart will be much higher than what was specified and the control limit will be larger than it should. This will make it more difficult for the chart to signal, and will make it harder to detect space-time interactions when they occur.

The changes in the simulated in-control ARLs for different space and time thresholds occur because the distribution of $z_{st}(i)$ changes for different thresholds. Some threshold combinations lead to more extreme departures from normality than others. To investigate the adequacy of the normal approximation of $z_{st}(i)$ for these thresholds, times and locations of disease occurrences were randomly generated for the same in-control situation used to investigate ARL performance in Section 2.3.1. A total of 100,500 observations were generated within a unit square region where the population density remained uniform and where the times between occurrences followed an exponential distribution with mean equal to 1. After each observation was generated, $z_{st}(i)$ was computed using the exact variance of $N_{st}(i)$ in equation (2.3). Each $z_{st}(i)$ value was calculated sequentially as would be done in the CUSUM monitoring scheme. This was repeated for each of the space and time threshold combinations that were used to determine ARL performance for the unit square uniform population density case.

To summarize the distribution of $z_{st}(i)$ for each threshold combination, the mean, variance, and standard error of the mean were estimated. When the number of observations is low, the value $n_s(i)$ is often zero. If $n_s(i)$ equals zero then the variance of $N_{st}(i)$ is also zero, making $z_{st}(i)$ impossible to compute. To avoid missing $z_{st}(i)$ values in the mean and variance computations, only the final 100,000 values of $z_{st}(i)$ were

used out of the total of 100,500 observations. The summary values of $z_{st}(i)$ for each threshold combination are shown in Table 2.4.

Table 2.4: Estimated means, variances, and standard errors of the standardized local Knox statistic for different space and time threshold combinations

s	t	*Mean*	*Variance*	*Std. Error*
0.1262	0.5	-2.494	0.577	0.0024
	1	-1.855	0.523	0.0023
	2	-1.415	0.507	0.0023
	6	-1.058	0.503	0.0022
0.1784	0.5	-1.891	0.554	0.0024
	1	-1.450	0.514	0.0023
	2	-1.165	0.514	0.0023
	6	-1.006	0.503	0.0022
0.2821	0.5	-1.389	0.519	0.0023
	1	-1.139	0.506	0.0022
	2	-1.021	0.503	0.0022
	6	-1.102	0.494	0.0022
0.3989	0.5	-1.175	0.509	0.0023
	1	-1.038	0.494	0.0022
	2	-1.020	0.494	0.0022
	6	-1.266	0.508	0.0023

Rogerson (2001) assumed that the standardized local Knox statistic has an approximate standard normal distribution. For each set of threshold values, however, the mean of $z_{st}(i)$ is substantially lower than zero and the variance is approximately 0.5. This indicates that the mixture distribution, with the probability mass function given in equation (2.1), is inappropriate for modelling $N_{st}(i)$ when it is monitored sequentially. Since the mean of this statistic is negative and the variance is smaller than one, the CUSUM values, C_i, are smaller than they would be if the assumption that the mean is zero and the variance is one were true. This makes it harder for the CUSUM chart to signal and explains why the true in-control ARLs are larger than the in-control ARLs expected under normality. We expect that the mean is less than what would be

expected under the hypergeometric model because under this model observations can be close in time to those earlier or later in time, but in prospective monitoring the last observed point only has the opportunity to be close to earlier points.

To get a clearer picture of the distribution of $z_{st}(i)$ and how it differs from the normal distribution, a plot of the first 1,000 $z_{st}(i)$ values used to compute the summary statistics in Table 2.4 and a normal probability plot of these values were examined for each threshold combination. Figure 2.6 shows these plots for two space and time threshold combinations. The threshold combinations chosen were $s = 0.1262$ and $t = 0.5$, and $s = 0.3989$ and $t = 6$. The first combination includes the smallest space and time thresholds and the second combination includes the largest thresholds. These were chosen because they represent the two most extreme cases.

It is clear that the distribution of the standardized local Knox statistic does not resemble a standard normal distribution for any of the threshold combinations based on the mean and variance results in Table 2.4, but when the threshold values for space and time are both large, the distribution of $z_{st}(i)$ is much closer to a normal distribution than when the threshold values are both small. This is apparent from the normal probability plots in Figure 2.6. When the threshold values are small, the distribution of $z_{st}(i)$ is highly skewed to the right. This is evident in plot (b) of Figure 2.6. There is also a large separation in these observations. This happens because $N_{st}(i)$ typically has a value of zero or one when the space and time thresholds are small with zero occurring much more frequently. When the threshold values are large, the distribution of $z_{st}(i)$ is much closer to a normal distribution because the count $N_{st}(i)$ typically takes on a wider range of values.

Our conclusion is that when using the standardized local Knox statistic to monitor space-time interaction with a CUSUM chart, it is not reasonable to assume that the statistics will have an approximate normal distribution. Because the mean and variance of $z_{st}(i)$ are not 0 and 1, respectively, the CUSUM chart used to monitor the standardized local Knox statistic, with values obtained from equation (2.8), is not appropriate. This problem could be corrected by using the simulated mean and variance values in place of 0 and 1. If this chart is used for monitoring after incorporating new mean and variance values, then the design should still be based on simulated ARL results and not on the normal theory results since $z_{st}(i)$ is often not close to being

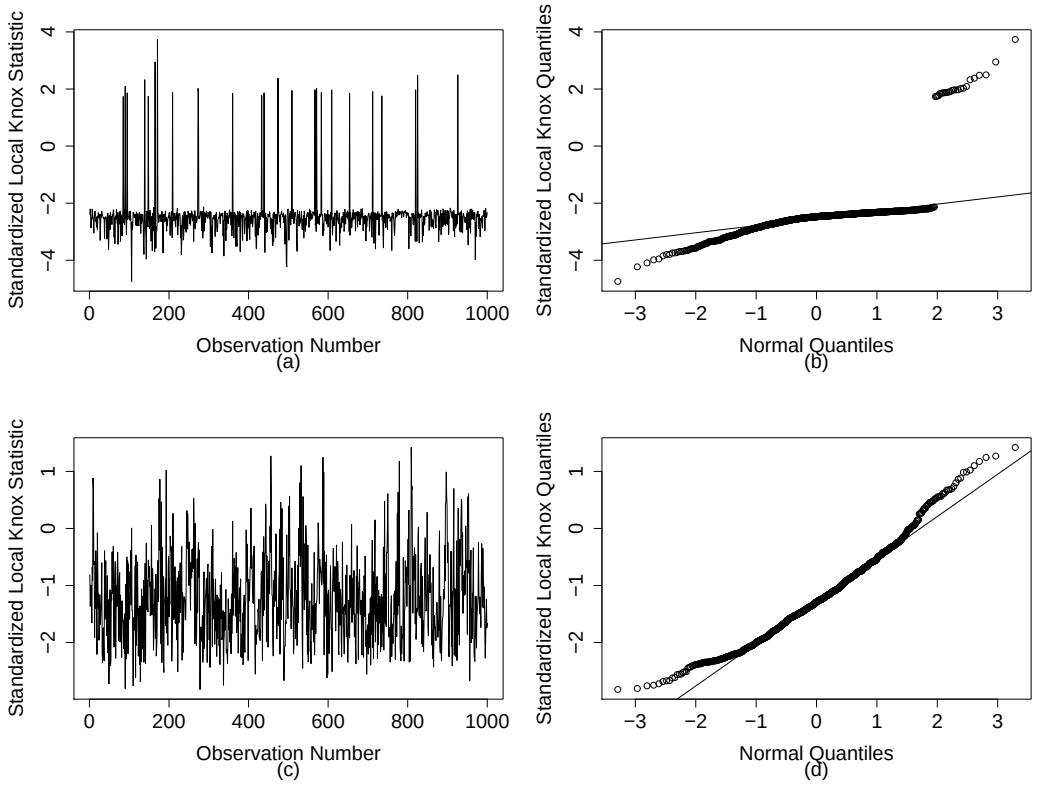


Figure 2.6: (a) Plot of 1,000 simulated $z_{st}(i)$ values when s=0.1262 and t=0.5; (b) Normal probability plot of 1,000 simulated $z_{st}(i)$ values when s=0.1262 and t=0.5; (c) Plot of 1,000 simulated $z_{st}(i)$ values when s=0.3989 and t=6; (d) Normal probability plot of 1,000 simulated $z_{st}(i)$ values when s=0.3989 and t=6

normally distributed, and because the space and time thresholds, population density, and region shape must be taken into account.

2.4 Summary and Discussion

The ARL performance of the CUSUM chart used to monitor the standardized local Knox statistic is influenced by many spatial and temporal elements. We considered the effect of the population density, changes in the population density over time, the region shape, the space and time thresholds, the value of the control limit, and different

variance expressions on the design of this chart. With the exception of the control limit, all of these elements change the distribution of $N_{st}(i)$ and, as a result, impact the in-control ARL performance of this monitoring method. Because the ARL is influenced by so many factors, it is difficult to determine the appropriate control limit for monitoring this statistic in any real-world setting. To design this chart the ARL performance must be determined through simulation for each new application.

Furthermore, there are issues dealing with the distribution of the the standardized local Knox statistic and the use of the CUSUM chart to monitor $z_{st}(i)$. The standard normal approximation for the local Knox statistic depends on the space and time thresholds and can be extremely poor. Therefore, it is not appropriate to use the in-control ARL values used for monitoring statistics that are normally distributed. This problem can be partially remedied by using mean and variance values obtained by simulation in place of the mean of zero and unit variance when calculating the CUSUM values, C_i. Another distributional issue stems from the fact that the distribution of $N_{st}(i)$ changes for each point. Some of these distributions may be highly skewed while others are more symmetric. The standardization of the statistic does not completely adjust for the differences between these distributions. The standardized statistics will have the same approximate mean and variance, but the shape of the distributions for each $z_{st}(i)$ can still be different.

In addition to the issues already discussed, there are some other issues that should be addressed if this method is to be implemented. These include the effect of the distribution of time between disease occurrences on in-control ARL performance, the impact of autocorrelation among the $z_{st}(i)$ values, which was mentioned by Lawson (2006), and the out-of-control ARL performance of the CUSUM chart. The out-of-control ARL performance determines the ability of this method to detect specified space-time interactions. A problem may arise when studying out-of-control ARL performance because changes in the population density will mimic the formation of disease clusters. If the control chart is designed to allow for changes in the population density over time without signaling, then it will also fail to signal when a disease cluster occurs. Therefore, this method may not be useful when the population density changes over the study period. Because of the complications addressed in regard to this method and these additional issues, the use of this method in any practical application can not be recommended.

Chapter 3

A Wavelet-Based Method for the Prospective Monitoring of Disease Occurrences

3.1 Introduction

The detection of adverse health events is important to reduce injury and maintain the health of the population. In particular, epidemiologists are concerned with the detection of clusters of disease and disease symptoms so that preventative measures can be taken in the hope of reducing morbidity and mortality. Researchers in the areas of epidemiology and statistics work to create surveillance systems for this purpose. These systems are large and have many stages. First, data on disease must be collected, which are typically in the form of disease incidence counts. Then these data are analyzed to identify possible cluster formation. Finally, the results must be put in the hands of a decision maker who will determine whether further investigation is necessary or whether action should be taken to contain the disease. Here we focus on the aspect of analyzing the data obtained through a surveillance system by developing a statistical method that can be used to determine if clusters of a specific disease are forming, and determine the location of these clusters in a specified geographical region.

There have been numerous methods developed for disease cluster detection, but the majority of them detect clusters in space or time, but not both. These methods are referred to as spatial and temporal surveillance methods, respectively. Methods that

detect clusters of disease in both space and time or detect clusters in space as they are forming in time are more rare. These methods are called spatio-temporal surveillance methods and are advantageous because of their ability to use information on both the time and location of disease incidences. Kulldorff (2001) developed a widely used scan method of this type and the wavelet-based surveillance method developed here is also a spatio-temporal method. Different spatio-temporal methods should be used depending on the particular application. One must decide between a retrospective or prospective method before designing or choosing a method, and it is also important to consider the type of disease under surveillance and the form of the available data. Diseases can be either chronic or infectious and data are either aggregated or have no aggregation. The wavelet-based method is a prospective method that can be used to monitor chronic diseases when aggregated data are available. An overview of the wavelet-based method is given in Section 3.1.1 and monitoring methods that share some similarities with the wavelet-based method are discussed in Section 3.1.2.

3.1.1 Wavelet-Based Surveillance Method Overview

The main goal of all spatio-temporal surveillance methods remains the same regardless of the type of analysis used, the type of disease monitored, or the type of data incorporated. This goal is to detect clusters of disease. The definition of a cluster, however, is not always the same for each method. In the wavelet-based surveillance method, a cluster is defined as an increase in the incidence rate per person in one subregion or more than one adjacent subregions within some geographical region. This method detects clusters based on this definition by prospectively monitoring an incidence rate surface modeled over the geographical region of interest. The incidence rate surface is modeled under the assumption that counts of incidences are observed for each subregion in the geographical region for equal time intervals, where the counts in each subregion are independent and the counts within each subregion follow a Poisson process over time. Therefore, this method should only be applied using aggregated space and time incidence count data that meet these assumptions. Use of the model is based on the assumption that all individuals within the region of interest have an independent risk of obtaining the disease. Therefore, this method should only be used to monitor chronic diseases.

An example of data appropriate for the wavelet-based method are the number of female respiratory lung cancer cases diagnosed in New Mexico from 1973 to 1991. These data include the number of female respiratory lung cancer diagnoses every month in each of the 32 counties located in New Mexico starting in the year 1973. In 1981, Valencia county was split into two counties, Valencia and Cibola, but the incidence counts are reported for these two counties as if they had remained one county. Yearly female population counts are also given along with the incidence counts. These data were collected through the Surveillance Epidemiology and End Results (SEER) program managed by the National Cancer Institute, and are available at www.seer.cancer.gov/resources.

The main challenges in creating a surveillance system for disease monitoring include the need to monitor a geographical region that is irregularly shaped, the need to detect both large and small clusters of disease, and the need to detect clusters that have different shapes. Wavelets can be used as a solution for these challenges and is the reason they are incorporated into this method. Wavelets meet the challenge of monitoring irregularly shaped regions by providing a way to map the region and its subregions to a grid. They also allow the mapped geographical region to be split into partitions for comparison. Wavelets are advantageous for detecting both small and large clusters because of their multiresolution. This provides the flexibility to model broad clusters that cover many subregions, as well as localized clusters that may only be present within one subregion. The multiresolution also leads to statistics with more power to detect clusters of a specified size. Wavelets can be used to detect clusters of different shape as well. Due to multiresolution, a cluster that includes any combination of adjacent subregions can be detected.

Once the geographical region is mapped to a grid, the process of cluster detection begins by modeling an incidence rate surface over the geographical region of interest each time new disease occurrence data are obtained. The surface is estimated using Poisson regression, where the regressors are functions from the Haar wavelet basis, the responses are disease incidence counts within each subregion, and population and covariate adjustments are made. It is important to note that using the Haar wavelets in the Poisson regression model is equivalent to the use of dummy variables. The advantage of using wavelets is that this parameterization emphasizes the proximity of subregions within a region, which is helpful for locating disease clusters.

After the surface has been estimated for a given time point, disease clusters are detected by using a control chart to monitor the model coefficients. This control chart is used to detect changes in the incidence rate surface. There are several control charts that can be selected for this purpose. The simplest is a multivariate chi-square control chart, which uses a Wald statistic to monitor the vector of model coefficients. A multivariate exponentially weighted moving average (MEWMA) control chart can also be used. These charts are known to be better at detecting smaller shifts in parameters than chi-square charts. Another alternative, is a control chart that uses a weighted version of the Wald statistic used in the chi-square chart, where the contribution of the model coefficients are given different weights. Both the chi-square and MEWMA control charts have difficulty detecting shifts in the parameter vector when there are a large number of parameters, which is the case in the wavelet-based method. The advantage of using a weighted statistic for monitoring is that it can alleviate this problem caused by high dimensionality.

When the control chart signals, diagnostics can be used to determine the resolution of the surface where incidence rates have changed. This resolution indicates the scope of a possible disease cluster. When the resolution is determined, the estimated increase in the disease incidence rate for the partitions of the region at this resolution can be used to determine cluster location. These values can be used to color code the original geographic map, which will show the subregions of increased incidence that are associated with a cluster. If more than one cluster is present, this will also be seen in the color-coded map.

The focus of the discussion on the wavelet-based surveillance method has been on the detection of clusters of disease, but the control chart used in this method is designed to detect any change in the incidence rate surface. Therefore, this chart can also detect decreased incidence rates of disease in the geographical region, which indicate negative clusters. In applications where it is important to detect both increases and decreases in incidence rates, this method is very useful. In applications where there is only interest in detecting disease clusters, this method is still very useful. In these cases, if the control chart signals only because the incidence rate in one or more subregions decreased, this will be apparent from the color-coded map because it will show no areas of highly elevated incidence rates. Therefore, no cluster will be indicated.

3.1.2 Similar Monitoring and Disease Surveillance Methods

There have been several methods developed for the prospective monitoring of disease clusters that share some of the same aspects of the wavelet-based method suggested here. The majority of the methods previously developed for prospective cluster detection are for cases when the available data are aggregated in both space and time, which is the type of data used in this method. The methods using data of this type are those of Raubertas (1989), Leung *et al.* (1999), Kulldorff (2001), Rogerson and Yamada (2004), Goovaerts and Jacquez (2005), Kleinman (2005), Sonesson (2007), Zhou and Lawson (2008), Neill (2009), and Neill and Cooper (2009). Of these methods, those of Leung *et al.* (1999), Kulldorff (2001), and Sonesson (2007) are the only methods similar to the wavelet-based method in the sense that they were developed specifically for monitoring chronic diseases, and therefore, do not incorporate spatial or temporal autocorrelation. The methods that incorporate spatial autocorrelation and are able to monitor infectious diseases, can also be used to monitor chronic diseases by replacing the spatial correlation matrix with an identity matrix. A component of each of these methods that is similar to the wavelet-based method, is that all of these methods are able to incorporate information on the population within each subregion, however, the methods of Raubertas (1989) and Leung *et al.* (1999) do not allow for the addition of any other covariates, in contrast to the wavelet-based method. Many of these methods use models that assume the underlying disease incidence counts come from a Poisson process. These methods include those of Raubertas (1989), Kulldorff (2001), Rogerson and Yamada (2004), Kleinman (2005), Sonesson (2007), Zhou and Lawson (2008), and Neill (2009). The method of Kleinman (2005) is the only one that uses Poisson regression to model these counts, as the wavelet-based method does. Instead of assuming a Poisson process, Goovaerts and Jacquez (2005) use a neutral spatial model based on randomization, where an assumption is made that all permutations of incidence rates in the subregions are equally likely for a given time. Leung *et al.* (1999) take a similar approach, but in this case an assumption is made that the permutations of incidence rates over time are equally likely within a subregion. To detect clusters of disease based on these models, Raubertas (1989), Leung *et al.* (1999), Rogerson and Yamada (2004), and Sonesson (2007) all use control charts. Rogerson and Yamada (2004) and Sonesson (2007) use a multivariate CUSUM control chart, which is similar to the wavelet-based

method, but Raubertas (1989) and Leung *et al.* (1999) use univariate CUSUM control charts for each subregion. Kulldorff (2001) calculates likelihood ratio statistics for multiple space-time windows over the region of interest and determines the subregions and time intervals corresponding to the likelihood ratio with the largest value. These subregions and time intervals constitute the most likely cluster. Goovaerts and Jacquez (2005), Kleinman (2005), and Zhou and Lawson (2008) detect clusters by calculating statistics for each subregion, which are used to create a color-coded map of the region indicating areas of increased disease incidence.

There are other disease surveillance methods that are similar to the wavelet-based method, but they are used in applications where there is no aggregation in the incidence data. These include the methods of Rogerson (2001) and Diggle (2005). The method of Diggle (2005) is similar to this wavelet-based method because it estimates a surface of incidence rate changes from a baseline over the geographical region. The method of Rogerson (2001) is similar to the wavelet-based method because it uses a control chart for detecting disease clusters.

None of the spatio-temporal methods for disease surveillance incorporate wavelets, but there have been two other applications of disease monitoring that use wavelets. These include the methods of Shmueli (2005) and Louie and Kolaczyk (2006). The method of Louie and Kolaczyk (2006) uses a "quad-tree", which is equivalent to using the Haar wavelet basis, to determine spatial clustering of disease at multiple resolutions. This method is also used for aggregated spatial data where a Poisson process is assumed, but it can not be used for detecting space-time clusters or for prospective monitoring, because it does not incorporate the time of disease incidences. Shmueli (2005) uses Haar wavelets to model and monitor highly autocorrelated syndromic surveillance data over time. This method is purely temporal and does not incorporate a spatial component.

The profile monitoring methods that have been developed in the area of statistical process control (SPC) are similar to the wavelet-based method and are discussed by Woodall *et al.* (2004). Profile monitoring is used to detect changes in the relationship between a response variable and one or multiple explanatory variables. In the wavelet-based method, the goal is to detect changes in the disease incidence rates, using explanatory variables that indicate the location of the incidences over a geographical

region. These explanatory variables are the Haar wavelet functions. Wavelets have previously been used for profile monitoring in SPC applications by Jeong *et al.* (2006), Zhou *et al.* (2006), and Reis and Saraiva (2006). In all of these cases, one dimensional profiles are monitored over time using a control chart. This application is different from these monitoring methods because a two-dimensional incidence rate surface is monitored over time.

3.1.3 Outline

Before beginning a formal discussion on the wavelet-based surveillance method, it is important to understand the underlying element of this method. Therefore, an introduction to wavelets is given in Section 3.2. In Section 3.3, each component of the method is described in detail. These components include the mapping of the geographical region to a grid, modeling the incidence rate surface, using a control chart to detect disease clusters, and determining the location and size of the clusters. In Section 3.4, several demonstrations of the wavelet-based method are shown using simulated data. These demonstrations show that the method is capable of detecting clusters of disease in different scenarios. Finally, in Section 3.5, a summary of the features of this method and its benefits is provided.

3.2 Introduction to Wavelets

In order to understand how wavelets are used in the surveillance method that has been developed, a brief overview of wavelets and their applications is given here. The information provided is only meant to aid in the comprehension of this method and should not be considered a complete treatment on wavelets. For more information on wavelets the reader is referred to Daubechies (1992) and Ogden (1997).

A wavelet family or wavelet basis is a set of wavelet functions that are similar in shape and form a complete orthonormal system in L_2-space. In other words, a wavelet basis is a set of similar orthonormal wavelets that can be used to approximate any square-integrable function $f : X \rightarrow \mathbb{R}$ arbitrarily well by taking a finite linear combination of the wavelet functions. Wavelets can be used to break down these functions

or signals into components of different scale or resolution. This decomposition of a function is similar to the Fourier decomposition. Here the lower resolution components represent the overall shape of the function or signal and the higher resolution components represent the more detailed aspects. The ability to look at these different layers of resolution is referred to as multiresolution analysis and is what makes wavelets useful in applications. An important distinction between the Fourier and wavelet decomposition of a function is that the wavelet decomposition uses a double-indexing scheme to indicate the resolution of a wavelet function in the multiresolution analysis, and this is not part of the Fourier analysis.

Wavelets have been used in many applications that are both statistical and non-statistical. Some common statistical applications involving wavelets include density estimation, Bayesian modeling, and nonparametric regression, which is the application used in the wavelet-based method. For more information on these techniques and others see Ogden (1997) and Vidakovic (1999). Some of the more traditional applications have been in signal processing, image processing, and data compression. The main use of wavelets in signal processing is to transmit an encoded signal and then reconstruct and denoise it once it has been received. In image processing, wavelets are also used to denoise. In these applications an imperfect image or picture is made clearer by denoising, while still maintaining the detail of the image. When wavelets are used for data compression the goal is to store data and images using a minimal amount of space. Wavelets are used to decompose these data or images. The components containing the important information are then stored and used to reconstruct the data when they are needed. For specifics on these conventional applications see Mallat (1999), Salomon (2000), and Chan and Shen (2005).

There are many wavelet bases to choose from in any application. Some common wavelet bases are the Haar basis and those in the system of Daubechies bases. Properties that are important to consider when selecting a basis are smoothness and the support of the wavelet functions. In many applications smoother wavelet functions give a better approximation at lower resolutions than wavelets that are not smooth. Compact support is important because typically a function is only examined over a bounded interval. The Haar basis has compact support, but it is not smooth. The Daubechies bases are smoother and also have compact support. Since the spatial component of the

data used for this application is aggregated, the smoothness of the wavelet functions is not a concern, and therefore, the Haar basis is used in this surveillance method because of its simplicity. Since the Haar basis is used in this application, the remainder of this discussion on wavelets will be presented in terms of the Haar wavelet basis.

Every wavelet basis is constructed from what is called the "mother wavelet" and has a scaling function, which is known as the "father wavelet". The mother wavelet for the Haar basis is a piecewise function on the interval from zero to one, which is given by

$$\psi(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} \leq x < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{3.1}$$

A plot of the Haar mother wavelet given in equation (3.1) is shown in Figure 3.1. Notice that this function takes on two values over the interval from zero to one. This splits the interval into two parts and this is how the Haar basis is used to subdivide the mapping of the geographical region into partitions in the wavelet-based surveillance method. The scaling function for the Haar basis is the indicator function on the interval from zero to one and is given by $\phi(x) = I_{[0,1)}(x)$.

The Haar wavelet basis can be constructed by performing two operations, dilation and translation, on the mother wavelet. By dilating the mother wavelet, the spread of the nonzero portion of the wavelet is either increased or decreased, and by translating the mother wavelet, the location of the wavelet is shifted to a different position on the real number line. Each wavelet in the Haar basis is represented by a function of the Haar mother wavelet given by

$$\psi_{j,k}(x) = 2^{j/2}\psi\left(2^j x - k\right), \tag{3.2}$$

where j is the dilation index, k is the translation index, and $j, k \in \mathbb{Z}$. It is important to

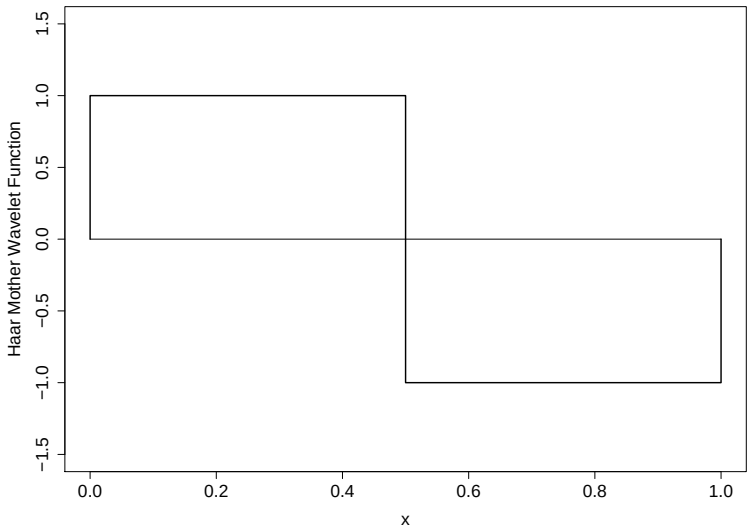


Figure 3.1: The Haar mother wavelet

note that the index j controls the resolution of the wavelet functions. As j increases, the spread of the nonzero portion of the wavelet functions decreases. This increase in resolution allows the features of smaller and smaller intervals to be approximated. The minimum and maximum values of j indicate the domain of the function under analysis and the index k, which depends on the values of j, indicates the location within the domain. The standard domain for the Haar wavelet basis is the interval [0,1). For this domain, $j = 0, 1, 2, \ldots$ and $k = 0, 1, 2, \ldots, 2^j - 1$. Some examples of wavelet functions from the Haar basis over this domain are shown in Figure 3.2. Each of these wavelet functions splits a portion of the interval into two parts. When the wavelet-based surveillance method is implemented, the mapping of the geographical region is subdivided into smaller and smaller halves as the resolution is increased and more of these functions are used.

While the Haar wavelet functions can be easily constructed from the mother wavelet, they are defined by linear combinations of dilations and translations of the Haar father wavelet. The dilations and translations of the Haar father wavelet are given by

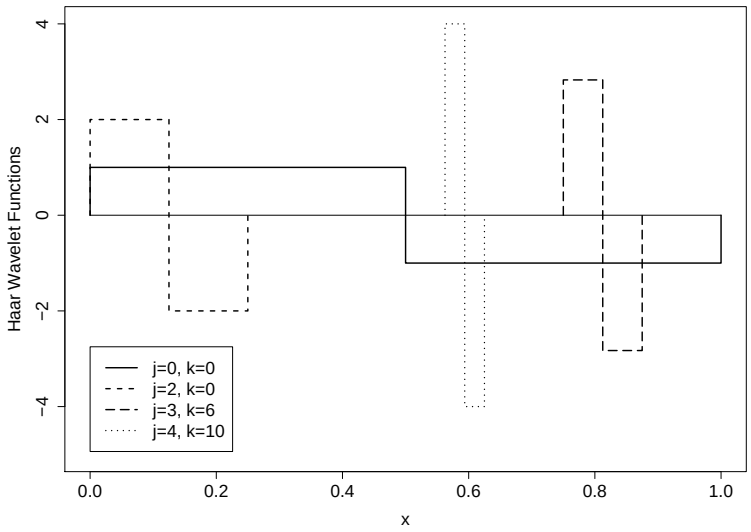


Figure 3.2: Dilations and translations of the Haar mother wavelet

$$\phi_{j,k}(x) = 2^{j/2} I_{\left[\frac{k}{2^j}, \frac{k+1}{2^j}\right)}(x), \tag{3.3}$$

where I is the indicator function on the interval from $k/2^j$ to $(k+1)/2^j$. The Haar wavelet functions are then defined as

$$\psi_{j,k}(x) = 2^{-1/2} \left[\phi_{j+1,2k}(x) - \phi_{j+1,2k+1}(x)\right]. \tag{3.4}$$

One of the main uses of wavelets is to approximate functions. As was mentioned earlier, a square-integrable function can be approximated by a finite linear combination of wavelet functions. If the function is approximated over the interval [0,1) using the Haar wavelet basis, then the approximation is

$$f(x) \approx a_0\phi(x) + \sum_{j=0}^{J}\sum_{k=0}^{2^j-1} a_{j,k}\psi_{j,k}(x), \tag{3.5}$$

where J is the value of the dilation index corresponding to the highest resolution, a_0 and $\{a_{j,k}|j = 0, 1, \ldots, J, k = 0, 1, \ldots, 2^j - 1\}$ are the wavelet coefficients, and $x \in [0, 1)$. The coefficient a_0 is the inner product of f and ϕ and the coefficients $\{a_{j,k}|j = 0, 1, \ldots, J, k = 0, 1, \ldots, 2^j - 1\}$ are the inner products of f and $\psi_{j,k}$, which are

$$a_0 = \int_0^1 f(x)\phi(x)dx \quad \text{and} \tag{3.6}$$

$$a_{j,k} = \int_0^1 f(x)\psi_{j,k}(x)dx, \tag{3.7}$$

respectively. The quality of this approximation improves as the value of J increases.

In many applications the underlying function of interest is unknown. In this case the Haar approximation in equation (3.5) can also be used in a regression setting to estimate a function from empirical data. In a case where there is only one independent variable of interest, one can simply calculate the Haar wavelet functions for the different values of the independent variable and use multiple linear regression to estimate the wavelet coefficients. To illustrate this, consider an example where the relationship between two variables is a cubic function, $f(x) = 20(x - 0.5)^3 + 0.5$. Suppose that random responses were obtained for different values of the independent variable on the interval $[0, 1)$ as shown in Figure 3.3 plot (a). To estimate the underlying function, a value for J must be selected and then the Haar wavelet functions, $\psi_{j,k}$ for $j = 0, 1, \ldots, J$ and $k = 0, 1, \ldots, 2^j - 1$ can be calculated for each value of the independent variable. These wavelet functions now become the regressors in a multiple linear regression model, where the estimated coefficients correspond to the wavelet coefficients. The estimated coefficients, a_0 and $\{a_{j,k}|j = 0, 1, \ldots, J, k = 0, 1, \ldots, 2^j - 1\}$, can then be used to obtain predictions. Plots of the predicted values for several values of J are given in Figure

3.3 plots (b)–(f). Notice that all of these estimates resemble histograms. This is due to the piecewise nature of the Haar basis. The lower resolution estimates show the general shape of the underlying function because the responses are essentially averaged over a larger area. As the resolution increases and this interval becomes smaller, the estimates begin to reproduce the empirical data instead of estimating the underlying curve. This is analogous to overfitting a regression model or choosing a bandwidth that is too small when using a smoothing method. Because of this issue, the selection of J is important in wavelet applications.

In some cases there may be more interest in estimating a two-dimensional surface rather than a curve, and that is the case for this surveillance method, where the goal is to estimate an incidence surface. In order to estimate a surface, a two-dimensional wavelet basis is needed. A two-dimensional Haar wavelet basis can be constructed from two one-dimensional Haar bases. This wavelet basis is formed from the one-dimensional wavelet functions for $x_1 \in [0, 1)$ and $x_2 \in [0, 1)$ and their crossproducts. This wavelet basis consists of the functions

$$\Psi_1(x_1, x_2) = \psi_{j_1,k_1}(x_1)\phi(x_2) \tag{3.8}$$

$$\Psi_2(x_1, x_2) = \phi(x_1)\psi_{j_2,k_2}(x_2) \tag{3.9}$$

$$\Psi_{1\times 2}(x_1, x_2) = \psi_{j_1,k_1}(x_1)\psi_{j_2,k_2}(x_2), \tag{3.10}$$

where $\Psi_1(x_1, x_2)$ are the wavelet functions for the first dimension, $\Psi_2(x_1, x_2)$ are the wavelet functions for the second dimension, and $\Psi_{1\times 2}(x_1, x_2)$ are their crossproducts. The indices $j_1, j_2, k_1, k_2 \in \mathbb{Z}$ are comparable to j and k in the one-dimensional case. Here j_1 and j_2 are the dilation indices for dimension one and dimension two, respectively, and k_1 and k_2 are the translation indices for dimension one and dimension two, respectively. Ogden (1997) describes a similar two-dimensional wavelet basis that only uses one dilation index as opposed to two, but for simplicity the two-dimensional basis in equation (3.8) is used in the wavlet-based disease surveillance method.

Just as the one-dimensional Haar wavelet basis can be used to approximate a function, the two-dimensional Haar wavelet basis can be used to approximate a square-

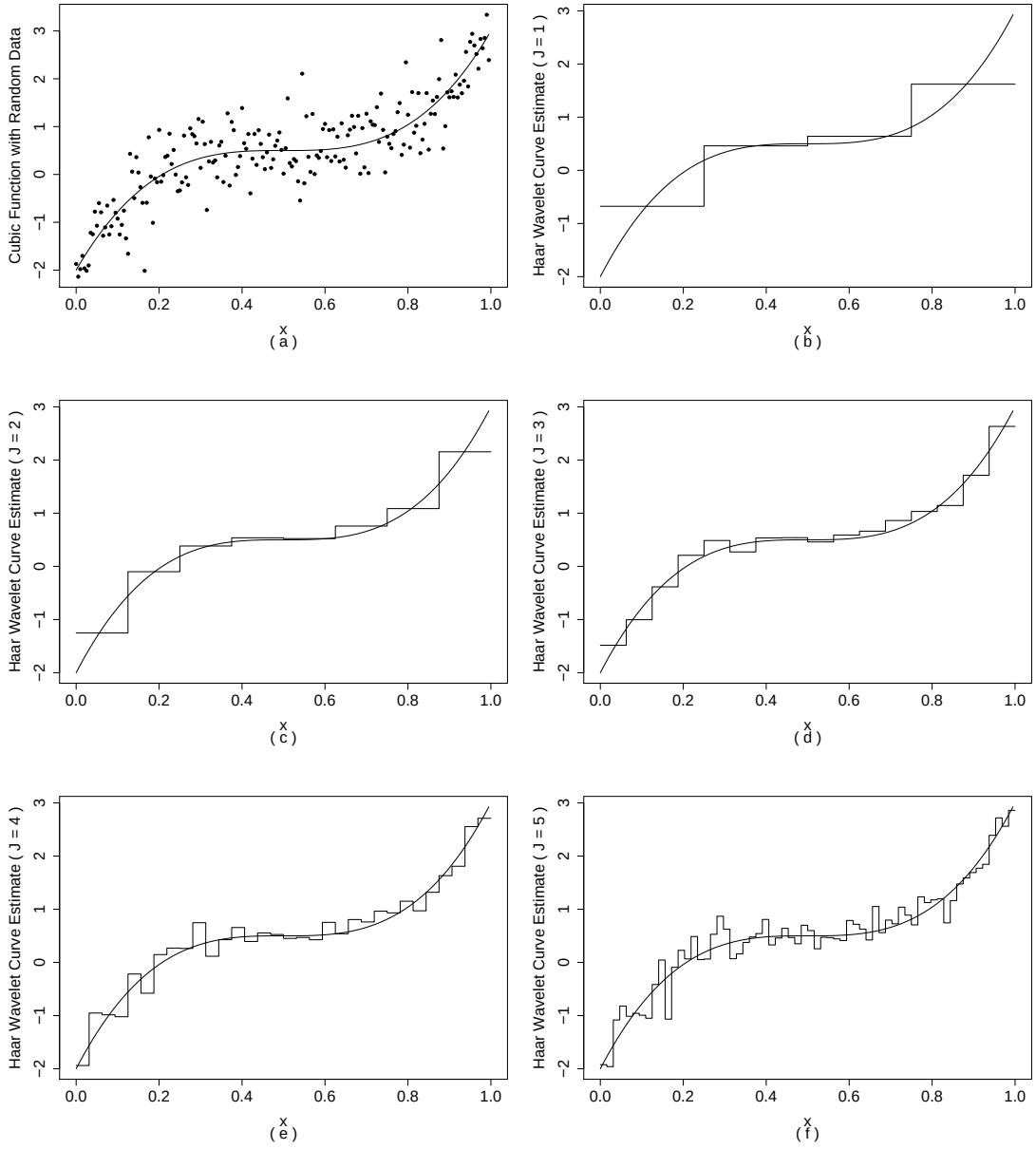


Figure 3.3: One-dimensional Haar wavelet basis estimates: (a) Underlying cubic function and empirical data on the interval $[0, 1)$; (b) Haar wavelet estimate for $J = 1$; (c) Haar wavelet estimate for $J = 2$; (d) Haar wavelet estimate for $J = 3$; (e) Haar wavelet estimate for $J = 4$; (f) Haar wavelet estimate for $J = 5$

integrable surface over the real plane, $f(x_1, x_2) \in L_2$. In this case the approximation function is

$$\begin{aligned} f(x_1,x_2) &\approx a_0\phi(x_1)\phi(x_2) \\ &+ \sum_{j_1=0}^{J_1}\sum_{k_1=0}^{2^{j_1}-1} a_{j_1,k_1}\psi_{j_1,k_1}(x_1) + \sum_{j_2=0}^{J_2}\sum_{k_2=0}^{2^{j_2}-1} a_{j_2,k_2}\psi_{j_2,k_2}(x_2) \\ &+ \sum_{j_1=0}^{J_1}\sum_{j_2=0}^{J_2}\sum_{k_1=0}^{2^{j_1}-1}\sum_{k_2=0}^{2^{j_2}-1} a_{j_1,k_1,j_2,k_2}\psi_{j_1,k_1}(x_1)\psi_{j_2,k_2}(x_2), \end{aligned} \tag{3.11}$$

where J_1 and J_2 are the values of the dilation indices corresponding to the highest resolution for the first and second dimension, respectively, and a_0, $\{a_{j_1,k_1}|j_1 = 0, 1, \ldots, J_1, k_1 = 0, 1, \ldots, 2^{j_1} - 1\}$, $\{a_{j_2,k_2}|j_2 = 0, 1, \ldots, J_2, k_2 = 0, 1, \ldots, 2^{j_2} - 1\}$, and $\{a_{j_1,k_1,j_2,k_2}|j_1 = 0, 1, \ldots, J_1, k_1 = 0, 1, \ldots, 2^{j_1} - 1, j_2 = 0, 1, \ldots, J_2, k_2 = 0, 1, \ldots, 2^{j_2} - 1\}$ are the wavelet coefficients.

To show how the two-dimensional Haar wavelet basis can be used to estimate a surface, a similar example to that given in Figure 3.3 is given in Figure 3.4. Here an assumption is made that data with normally distributed random errors were observed from the underlying surface shown in Figure 3.4 plot (a). Then the approximation function in equation (3.11) was used to obtain estimates of the surface for $J = J_1 = J_2 \in \{0, 1, 2, 3, 4\}$, which are shown in Figure 3.4 plots (b)–(f). As the resolution increases, these estimates look more like the underlying surface but they still resemble a histogram because of the shape of the mother wavelet.

The estimation of the incidence rate surface in the wavelet-based surveillance method is very similar to the estimation of the cubic surface shown in Figure 3.4. The main differences are that when estimating the incidence rate surface, the data are aggregated and the coefficient estimates are based on a Poisson model. When the data used to estimate a surface are aggregated, the highest resolution that can be used is bounded by the level of aggregation. Therefore, it is typical for the incidence rate surface estimates to resemble plots (c) and (d) of Figure 3.4 and not appear as smooth as plots (e) and (f) of Figure 3.4, because surface estimates at resolutions as high as $J = 3$ and $J = 4$ can not be obtained.

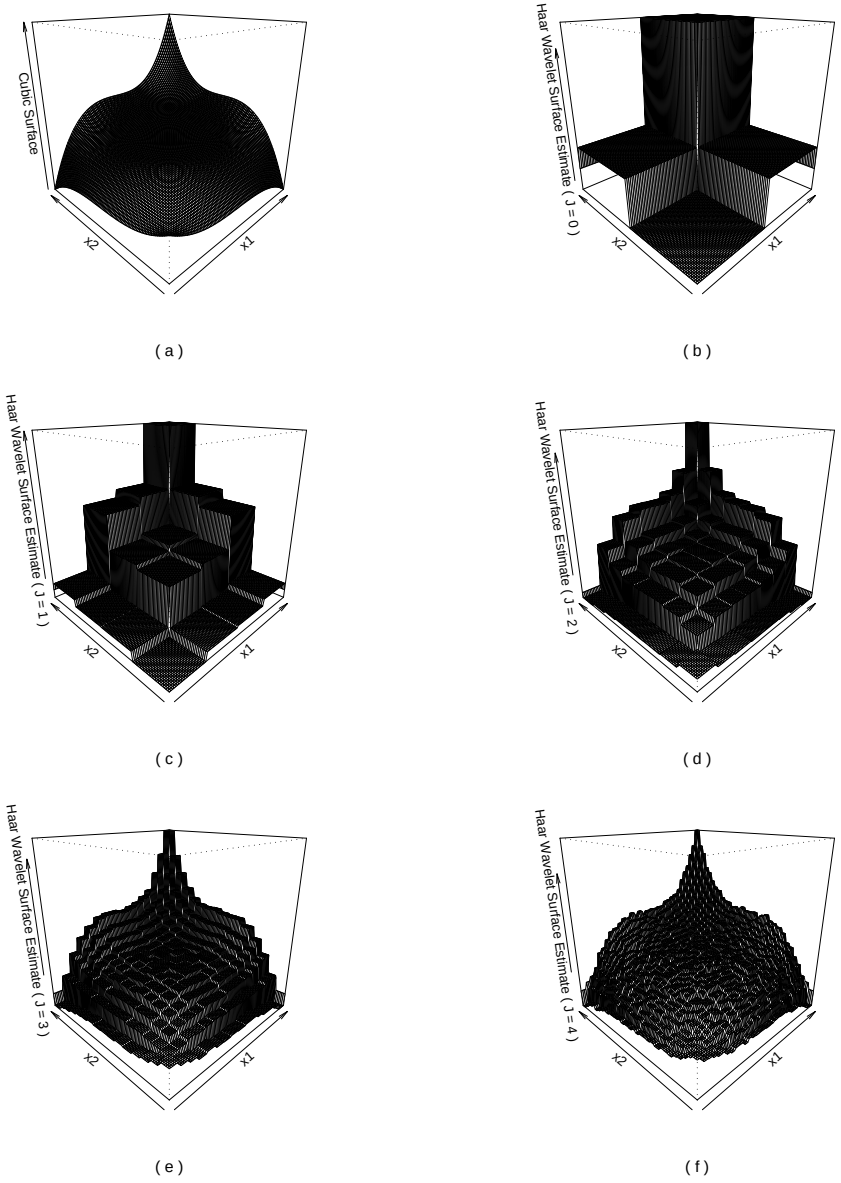


Figure 3.4: Two-dimensional Haar wavelet basis estimates: (a) Underlying cubic surface where $x_1 \in [0, 1)$ and $x_2 \in [0, 1)$; (b) Haar wavelet estimate for $J = 0$; (c) Haar wavelet estimate for $J = 1$; (d) Haar wavelet estimate for $J = 2$; (e) Haar wavelet estimate for $J = 3$; (f) Haar wavelet estimate for $J = 4$

3.3 Wavelet-Based Surveillance Method

The implementation of the wavelet-based monitoring method requires the mapping of the geographical region to a grid in the wavelet domain, the modeling of a surface of incidence rates over the region of interest, and the monitoring of this surface over time to determine if it is changing from a known baseline. When a change has been detected in the incidence surface, the location of the change must also be determined. Each of these aspects of the wavelet-based surveillance method is described in this section. The mapping of the region is discussed in Section 3.3.1 and the estimation of the surface is explained in Section 3.3.2. Surface monitoring is discussed in Section 3.3.3 and methods for determining cluster location are covered in Section 3.3.4.

3.3.1 Mapping of the Geographical Region

Obtaining an estimated incidence rate surface using the two-dimensional Haar wavelet basis requires the mapping of the subregions in the geographical region of interest to the wavelet domain. This domain is a grid of square or rectangular cells of equal area, where the number of grid cells, denoted by N_{wd}, is equal to a power of two. The number of grid cells in this domain is equal to a power of two as a result of the wavelet functions splitting partitions of their domain in half each time the resolution is increased. An example of the mapping of subregions of a geographical region to the wavelet domain is shown in Figure 3.5. This is one possible mapping of the counties in the state of New Mexico, which could be used to monitor female respiratory cancer incidences.

The power of two that determines the number of grid cells in the wavelet domain is based on the number of subregions in the geographical region, which is denoted by N_{gr}. If the number of geographical subregions is equal to a power of two, then the number of grid cells will equal the number of geographical subregions, making $N_{wd} = N_{gr}$. This will not typically be the case, because it is unlikely that a region will have a number of subregions equal to an exact power of two. When the number of subregions is not equal to a power of two, then some subregions are pooled so that their number is reduced to a power of two. In this case, the number of grid cells in the wavelet domain is $N_{wd} = 2^r$, where r is a nonnegative integer chosen so that $2^r < N_{gr}$. The subregions are pooled by an iterative algorithm that combines the subregion of lowest population

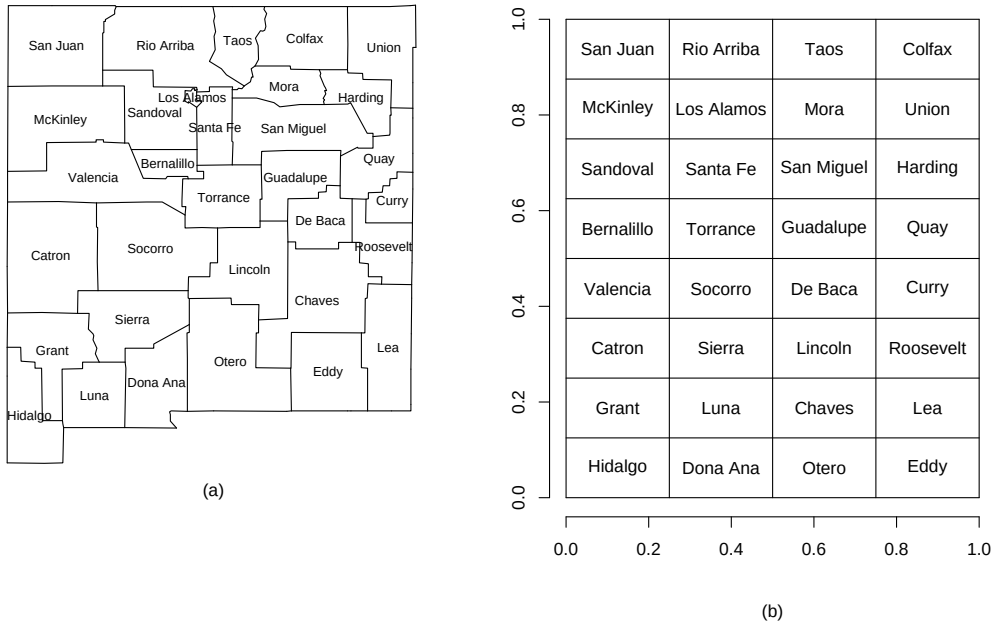


Figure 3.5: Wavelet domain mapping example: (a) Counties in New Mexico; (b) Possible mapping of counties to the two dimensional Haar wavelet domain

with its adjacent subregion of lowest population repeatedly, until the number of pooled subregions reaches $N_{wd} = 2^r$. These N_{wd} pooled subregions are then referred to as the subregions of the geographical region.

There is an alternative to pooling which uses 2^{r+1} grid cells instead of 2^r cells. In this case, each subregion would correspond to one grid cell leaving $2^{r+1} - N_{gr}$ empty cells. To implement the surveillance method, these $2^{r+1} - N_{gr}$ cells would be filled with fabricated subregions with population size equal to zero. This approach is not used for several reasons. One problem is that the number of unused grid cells can be very large. If this occurs then there may be large areas in the wavelet domain with no population, making

it difficult for the wavelet functions to compare different partitions of the domain. Another problem arises in cases where a rare disease is being monitored. In this situation, the incidence counts will be low, especially in subregions with low population. When these counts are low, coefficient estimates from the Poisson regression model, used to estimate the incidence rate surface, can be unstable. Therefore, the pooling of subregions with the lowest populations is advantageous because it will increase the counts for these areas. An increase in the counts also improves the approximate distribution of the statistics used in the control chart for detecting changes in the incidence rates. Another advantage of pooling over using 2^{r+1} grid cells, is the reduction in the number of wavelet functions needed to partition the wavelet domain. This also reduces the number of coefficients in the incidence surface model and the number of coefficients monitored with the control chart. Monitoring fewer coefficients will improve the efficiency of the control chart to detect clusters of disease.

Once the number of grid cells is determined, the resolution for each dimension of the wavelet domain and an optimal assignment of the geographical subregions within the wavelet domain must be established. The resolutions determine the dimensions of the grid within the two-dimensional wavelet domain. The resolution for dimension 1 and dimension 2 are denoted J_1 and J_2, respectively. In order to obtain $N_{wd} = 2^r$ grid cells, $J_1 + J_2 + 2$ must equal r. Therefore, there are a finite number of resolution pairs, which are $\{(J_1, J_2)|J_1 = 0, 1, ..., (r-2); J_2 = (r-2) - J_1\}$. Notice that the pairs $(0, r-2)$ and $(r-1, 0)$, $(1, r-3)$ and $(r-3, 1)$, ..., $([(r-2)/2], (r-2)-[(r-2)/2])$ and $((r-2)-[(r-2)/2], [(r-2)/2])$, where $[.]$ is the greatest integer function, are equivalent. Therefore, only one pair from each group needs to be considered. The dimension resolutions and assignments of the geographical subregions are chosen to optimize an objective function that evaluates the similarity of the locations of the subregions in the wavelet domain compared to their locations in the geographical map. Several choices for this objective function are discussed in Section 3.3.1.1 and are evaluated in Section 3.3.1.2. Once an objective function is selected, the optimal mapping is found by using an algorithm that is able to determine the wavelet domain assignment that optimizes the value of the objective function over all possible assignment permutations for each resolution pair. The optimal mapping found can also be transposed or flipped if the positions of the subregions in the transposed or flipped mapping match the true map

more closely. Transposing or flipping the optimal mapping merely switches the grid dimensions, thus, it is equivalent.

There will be cases when there is not a unique optimal mapping for the objective function. When this happens, all optimal mappings should be considered based on some other criterion. For example, there may be cases when more importance should be placed on particular subregions. This may be because they have large populations or because there is more concern that the incidence rate in these subregions and surrounding subregions will change. In these cases, the location of the most important subregions and their surrounding subregions should be considered when selecting an optimal mapping.

There may also be additional restrictions to consider when determining a mapping of a geographical region to the wavelet domain. For instance, sometimes a certain area of the geographical region may be of particular interest. Then it is important that the subregions within this area maintain adjacency in the wavelet domain. In this case, an optimal layout should be chosen from the subset of all possible layouts where the subregions within this area are adjacent. This can be done by imposing a restriction to only consider assignment permutations from this subset.

3.3.1.1 Objective Functions

When implementing the permutation algorithm to determine an optimal assignment of subregions to the wavelet domain, it is important to choose an objective function that will find a mapping that maintains the location characteristics of the geographical region. Three objective functions are presented here with the aim of meeting this criterion. These objective functions compare the distance, direction, and adjacency of each pair of subregions within the wavelet domain to their true distance, direction, and adjacency in the geographical region.

The distance objective function evaluates the difference in the distances between each pair of subregions in the wavelet domain mapping and the geographical map. This objective function is

$$\text{ObjFn(DIST)} \quad = \quad \sum_{i=1}^{N_{wd}-1} \sum_{j=i+1}^{N_{wd}} ABS\left(D(wd)_{ij} - D(gr)_{ij}\right), \tag{3.12}$$

where $D(wd)_{ij}$ is the Euclidean distance between the centroids of subregions i and j in the wavelet domain, $D(gr)_{ij}$ is the scaled Euclidean distance between population centers, such as capitals or county seats, of subregions i and j in the geographical region, and N_{wd} is the number of subregions in the wavelet domain. The geographical distances $D(gr)_{ij}$ are scaled so that their range is equivalent to the range of distances between centroids of subregions in the wavelet domain. This allows for the comparison of geographical and wavelet domain distances. In cases where a subregion is comprised of two or more pooled areas, the population center is determined by taking the midpoint between the population centers of the areas as they are pooled. When the distances between each pair of subregions in the wavelet domain mapping are similar to the scaled distances in the geographical region, this objective function should have a small value. Therefore, when using the distance objective function to determine an optimal mapping, the assignment of subregions to the wavelet domain that achieves the minimum value of this objective function should be chosen.

Another objective function that can be used to determine an optimal mapping is one that considers the relative direction of each pair of subregions in the wavelet assignment to their relative direction in the geographical region. An objective function used to evaluate relative direction is

$$\text{ObjFn(DIR)} \quad = \quad \sum_{i=1}^{N_{wd}-1} \sum_{j=i+1}^{N_{wd}} D_{ij}, \tag{3.13}$$

where

$$D_{ij} = \begin{cases} 1 & \text{if subregions } i \text{ and } j \text{ are in the same relative direction} \\ & \text{in the wavelet domain and in the geographical map} \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

and N_{wd} is the number of subregions in the wavelet domain. To determine the relative direction of two subregions in the wavelet domain, the centroid of one subregion, say

subregion 1, is assumed to be the origin and the angle of a ray that starts at the origin and passes through another subregion, say subregion 2, is calculated. Using this angle, the direction of subregion 2 in relation to subregion 1 is determined according to Table 3.1. The relative direction of two subregions in the geographical region is determined in the same manner as in the wavelet domain with the exception that the ray starts at the population center of subregion 1 and passes through the population center of subregion 2 instead of passing through the centroids. The population centers are determined using the same approach used for the distance objective function. Once the relative directions of the subregions are determined in the wavelet domain and in the geographical region, they are compared to calculate the objective function. In this case a large value of the objective function is preferred because this indicates that more pairs of subregions are in the same relative directions in the wavelet map when compared to the geographical map. The optimal wavelet mapping in this case would be the assignment that maximizes the objective function.

Table 3.1: Relative subregion directions for calculated angles

Angle		
Degrees	Radians	Direction
$-45° < \angle < 45°$	$-\frac{\pi}{4} < \angle < \frac{\pi}{4}$	East
$\angle = 45°$	$\angle = \frac{\pi}{4}$	Northeast
$45° < \angle < 135°$	$\frac{\pi}{4} < \angle < \frac{3\pi}{4}$	North
$\angle = 135°$	$\angle = \frac{3\pi}{4}$	Northwest
$135° < \angle < 225°$	$\frac{3\pi}{4} < \angle < \frac{5\pi}{4}$	West
$\angle = 225°$	$\angle = \frac{5\pi}{4}$	Southwest
$225° < \angle < 315°$	$\frac{5\pi}{4} < \angle < \frac{7\pi}{4}$	South
$\angle = 315°$	$\angle = \frac{7\pi}{4}$	Southeast

It is important to point out that there is a caveat associated with this scheme for calculating a direction objective function. Since the wavelet domain is on a grid, the relative direction of several subregion pairs in the wavelet mapping will be Northeast, Northwest, Southwest, or Southeast. It is unlikely, however, that the relative direction of subregions in the geographical map will have these directions because their angles will not equal $\frac{\pi}{4}$, $\frac{3\pi}{4}$, $\frac{5\pi}{4}$, or $\frac{7\pi}{4}$, respectively. To account for this issue, if the relative direction of two subregions is classified as Northeast, Northwest, Southwest, or Southeast, it is

considered both North and East, North and West, South and West, or South and East, respectively. To see how this affects the calculation of D_{ij} more fully, suppose that the relative direction of a pair of subregions is North in the the geographical region. If the relative direction for this same pair of subregions is Northeast, North, or Northwest in the wavelet mapping, then D_{ij} would have a value of one. Otherwise, D_{ij} would have a value of zero.

A third objective function that can be used in the algorithm to determine an optimal wavelet domain mapping is one that considers the adjacency of subregions. The objective function for this case is

$$\text{ObjFn(ADJ)} = \sum_{i=1}^{N_{wd}-1} \sum_{j=i+1}^{N_{wd}} A_{ij}, \tag{3.15}$$

where

$$A_{ij} = \begin{cases} 1 & \text{if subregions } i \text{ and } j \text{ are adjacent in both} \\ & \text{the wavelet domain and the geographical map} \\ & \text{OR} \\ & \text{if subregions } i \text{ and } j \text{ are incontiguous in both} \\ & \text{the wavelet domain and the geographical map} \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

and N_{wd} is the number of subregions in the wavelet domain. A pair of subregions is considered adjacent in the wavelet mapping if their grid cells share a border. Subregions with grid cells that touch at only the corners are not considered adjacent. In the geographical map, subregions that share a portion of their borders are considered adjacent. As with the wavelet map, if two subregions touch only at a corner, they are not considered adjacent. After determining the adjacency of subregion pairs in the wavelet mapping and the geographical map, the objective function can be calculated. As with the direction objective function, large values of the adjacent objective function

are favored. As a result, the optimal wavelet mapping should be the assignment of subregions that gives the maximum value for this objective function.

3.3.1.2 Use of a Random Search Algorithm

The mapping algorithm, which considers all possible permutations of subregion assignments to the wavelet domain, will produce an optimal mapping based on a chosen objective function, but in cases where the number of subregions is large, this can be too computationally intensive. When this occurs, the implementation of a random search algorithm is suggested to find a near-optimal mapping, or in some cases an optimal mapping, more quickly. A random search algorithm should be used when the number of subregions is $2^4 = 16$ or larger because the use of the permutation algorithm is prohibitive for these cases.

When it is necessary to use a random search algorithm to determine a mapping, one common search algorithm that can be implemented is a simple local search. A simple local search starts by randomly selecting a solution within the solution space. In this application, a random assignment of subregions to the wavelet domain is obtained from all possible assignments, given that the dimensions of the wavelet domain have been determined. Then the algorithm evaluates this solution based on the objective function and also evaluates a solution that is randomly selected from the local neighborhood of the original solution. The solution that gives the most optimal value of the objective function is then chosen. For this application a local neighborhood is all assignments to the wavelet domain that differ by at most κ grid cells from the original assignment. The value of κ, which can be any integer from 2 to N_{wd}, is selected by the user. A reasonable approach for choosing κ is to start with $\kappa = 2$ and then to increase the value if the procedure fails to produce reasonable near-optimal solutions. The simple local search algorithm then continues until a stopping criterion is met. The stopping criterion is designed to indicate when the random search algorithm has converged to a near-optimal solution, which implies that going through more iterations of the algorithm is unnecessary. The procedure for using this simple local search algorithm to determine a wavelet domain mapping is outlined explicitly in STEPS 1 through 5 of Table 3.2.

A common problem that arises when using simple local search algorithms is that they sometimes converge to a local optimum as opposed to a global optimum. One way

to deal with this issue is to repeat the search algorithm using many randomly selected starting solutions. Then the near-optimal solution can be selected from the solutions produced by each random starting solution. This is done by choosing the solution that has the most optimal value of the objective function. When several random starting solutions are used, STEPS 1 through 6 in Table 3.2 should be followed for determining a near-optimal assignment of subregions to the wavelet domain. Using many random starting assignments is necessary to find a reasonable near-optimal assignment in this application.

Table 3.2: Simple random search algorithm for determining a near-optimal wavelet domain mapping of a geographical region

STEP 1:	Produce a random assignment of subregions to the wavelet domain.
STEP 2:	Calculate the value of the chosen objective function for the random assignment.
STEP 3:	Randomly switch the locations of at least κ subregions in the wavelet domain assignment and calculate the objective function for this new assignment.
STEP 4:	Compare the values of the objective function for the two wavelet domain assignments and keep the assignment that gives the optimal value of the objective function.
STEP 5:	If the stopping criterion is met, then STOP and keep the assignment chosen in STEP 4. Otherwise, GO TO STEP 3.
STEP 6:	Repeat STEPS 1 through 5 for a chosen number of random starting assignments. Then select the wavelet domain assignment from the assignments produced by each random starting solution that has the optimal value of the objective function. The wavelet domain assignment selected in this step is the near-optimal mapping.

To evaluate the performance of the simple local search algorithm and the objective functions in Section 3.3.1.1 for finding a near-optimal wavelet mapping of a geographical region, ten states in the United States (US) were selected to map to the wavelet domain. These states included Rhode Island, Connecticut, New Hampshire, Arizona, Maine,

Maryland, New Mexico, California, Louisiana, and Ohio. These states were chosen because the number of subregions, which are counties in this case, within these states cover a broad range and because their shapes vary widely.

Before the random search algorithm could be implemented for the ten states, the original counties were pooled so that the number of subregions equaled a power of two. The counties were pooled based on the population sizes of each county, which were obtained from the U.S. Census Bureau website, www.census.gov. Then the resolution for each dimension of the wavelet domain was determined. The dimension resolutions for each state were chosen so that the resolutions were equal if the sum of the resolutions required to achieve the correct number of subregions was even or were chosen so that the resolutions had a difference of one if the sum of the resolutions required was odd. For example, if a state had 16 subregions, the resolution of each dimension of the wavelet domain was assigned one. If a state had 32 subregions, one dimension was assigned a resolution of one and the other was assigned a resolution of two. This produced a grid in the wavelet domain that allowed for the largest number of adjacent subregions.

Once the counties were pooled and the wavelet domain dimensions were determined, the simple random search algorithm outlined in Table 3.2 was implemented for each state using each objective function. The value of κ used was 2 and the stopping criterion was defined as

$$I_{\text{current}} - I_1 \quad > \quad 10 \left[\frac{\sum_{i=1}^{3} (I_i - I_{i-1})}{3} \right], \tag{3.17}$$

where I_{current} is the current iteration number, I_1 is the iteration where the last assignment change occurred, and I_i, for $i = 2, 3, 4$, is the iteration where the i^{th} to last assignment change occurred. By using this rule, the algorithm stopped when the number of iterations since the last change in the assignment became greater than ten times the average number of iterations needed to produce the last three assignment changes. The simple local search algorithm was repeated for 100,000 random starting assignments and the wavelet mappings with the most optimal values were selected for each state and objective function. In cases where there was not a unique wavelet domain

assignment for the most optimal value of the objective function, the wavelet mapping that most closely matched the geographical region or the mapping that was most representative of all the solutions was selected. The wavelet domain mappings for each state and objective function are presented in Appendix C, along with the original and pooled state maps.

The wavelet domain mappings for the ten states produced using the search algorithm outlined in Table 3.2, in conjunction with the stopping rule given in equation (3.17) and the objective functions in equations (3.12), (3.13), and (3.15), show that no single objective function outperforms the others. For Rhode Island and Connecticut, all of the objective functions were able to produce mappings that reasonably represented the true geographical map. Maine was the only state where reasonable mappings were given by two objective functions, which were the distance and direction functions. New Hampshire, Arizona, Maryland, Louisiana, and Ohio each had only one objective function that gave a reasonable mapping. For Arizona and Maryland, this was the distance objective function, for Louisiana and Ohio, this was the direction objective function, and for New Hampshire, this was the adjacent objective function. None of the objective functions produced a reasonable mapping for New Mexico or California.

When the simple local search failed to find a reasonable mapping, there were two main problems that tended to occur. The first problem was that the algorithm split the state into two subgroups in order to optimize the chosen objective function. This tended to happen in cases where the geographical orientation of the subregions was far from grid-like. The two ways the algorithm typically split the states are represented in Figure 3.6 below. This problem occurred for the States of New Hampshire, New Mexico, and California with both the distance and direction objective functions. It also occurred for the State of Arizona with the adjacent objective function.

The second problem that occurred was that in some cases the algorithm would misplace a single subregion or a small set of subregions to align the majority of the subregions as they appear in the geographical region. This tended to occur in cases where a subregion protruded from the edge of the state or in cases where the geographical orientation of most of the subregions aligned well to a grid if a few subregions were removed. An illustration of the typical cases when this problem arose are shown in Figure 3.7. This problem occurred for Louisiana and Ohio when using the distance

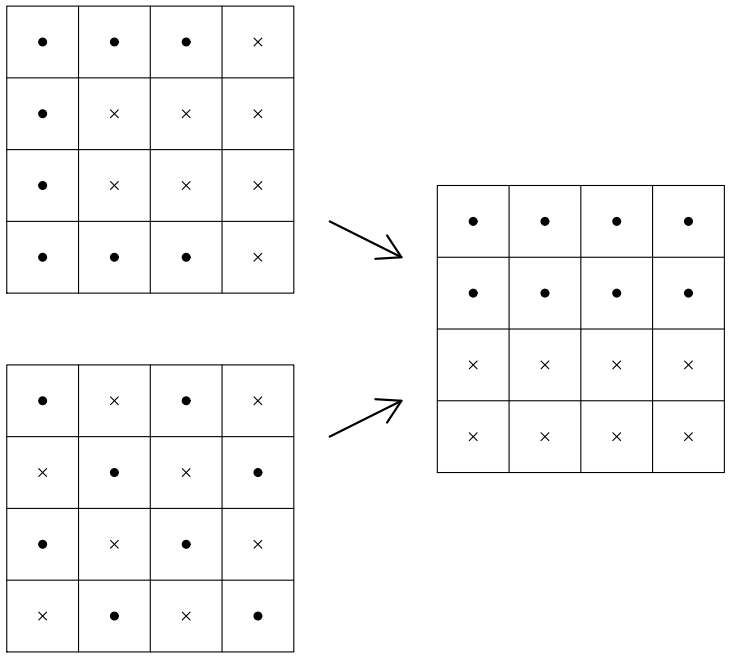


Figure 3.6: Illustration of mapping search algorithm problem 1

objective function, for Arizona and Maryland when using the direction objective function, and for Maine, Maryland, New Mexico, California, Louisiana, and Ohio when using the adjacent objective function.

A summary of the mapping results and the issues that arose for each state and objective function is shown in Table 3.3. In addition to evaluating the search algorithm based on this summary, the search algorithm must also be evaluated based on the number of random starting assignments needed to produce results similar to those in Appendix C. To determine the estimated number of random starting assignments necessary to produce results similar to those found for the ten states examined, the probability of the algorithm producing the mappings found in Appendix C was esti-

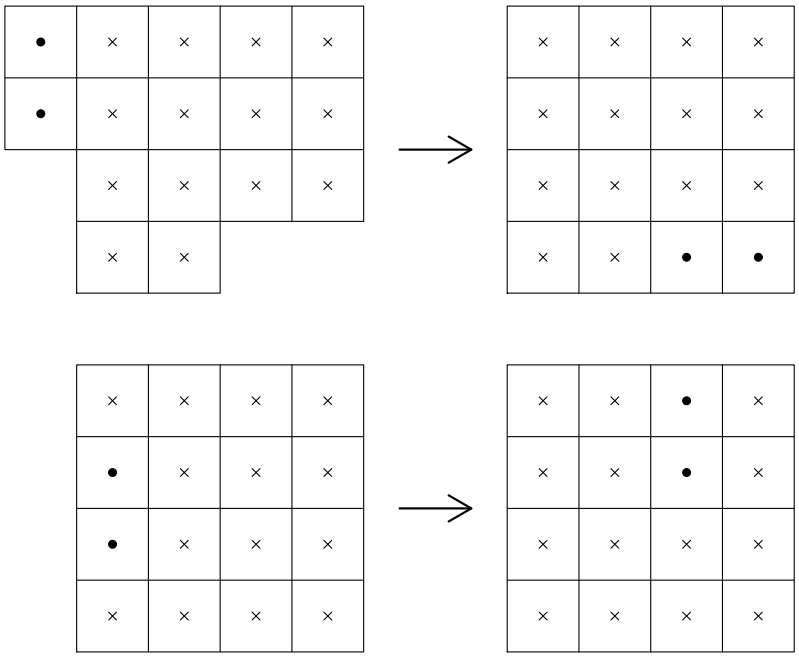


Figure 3.7: Illustration of mapping search algorithm problem 2

mated from the 100,000 runs of the algorithm for each state and objective function. Then this probability was used to estimate the number of random starting assignments needed to obtain these maps with 90%, 95%, and 99% certainty. The formula used to estimate the number of random starting assignments needed was

$$\text{P(Near-optimal mapping is found at least once)} = 1 - (1 - \hat{p})^n, \quad (3.18)$$

where $\hat{p}$ is the probability of the simple local search algorithm producing the near-

optimal state mapping for a single random starting assignment and n is the estimated number of random starting assignments needed to achieve a specified probability of finding the near-optimal mapping at least once. The estimated number of random starting assignments necessary to produce the maps for each state and objective function are in Table 3.4.

Table 3.3: Summary of mapping algorithm results for ten states in the US

State	Number of Subregions	Objective Function		
		Distance	Direction	Adjacent
Rhode Island	4	✔	✔	✔
Connecticut	8	✔	✔	✔
New Hampshire	8	$✗_1$	$✗_1$	✔
Arizona	8	✔	$✗_2$	$✗_1$
Maine	16	✔	✔	$✗_2$
Maryland	16	✔	$✗_2$	$✗_2$
New Mexico	32	$✗_1$	$✗_1$	$✗_2$
California	32	$✗_1$	$✗_1$	$✗_2$
Louisiana	64	$✗_2$	✔	$✗_2$
Ohio	64	$✗_2$	✔	$✗_2$

✔ = Algorithm produced a reasonable mapping
$✗_1$ = Algorithm produced a mapping with problem 1
$✗_2$ = Algorithm produced a mapping with problem 2

Table 3.4: Estimated number of random starting assignments for ten states in the US for each objective function

State	Number of Subregions	Objective Function								
		Distance			Direction			Adjacent		
		0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
Rhode Island	4	1	1	1	4	5	7	1	1	1
Connecticut	8	4	6	8	2	2	3	51	66	101
New Hampshire	8	30	40	60	13	17	25	39	50	77
Arizona	8	6	7	11	7	9	13	79	103	157
Maine	16	24	32	48	19	24	37	6,059	7,883	12,117
Maryland	16	16	20	31	31	40	61	16,446	21,397	32,892
New Mexico	32	328	427	656	221	287	441	115,129	149,786	230,257
California	32	710	924	1,420	223	290	445	230,258	299,572	460,515
Louisiana	64	23,025	29,956	46,050	1,145	1,489	2,289	230,258	299,572	460,515
Ohio	64	23,025	29,956	46,050	2,151	2,799	4,302	230,258	299,572	460,515

While the results vary from state to state, in general, the direction function seems to require the fewest number of random starting assignments to produce a near-optimal mapping. The distance objective function appears to require more than the direction objective function, but the adjacent objective function requires many more random starting assignments than either the distance or direction functions. For all objective functions, as the number of subregions per state increases, the number of random starting assignments needed to produce a near-optimal mapping increases rapidly. Based on the estimated number of random starting assignments needed to produce mappings for these ten states, it seems that either the direction or distance objective function would be preferable to use in the algorithm rather than the adjacent objective function. Although, while these results are helpful in determining how quickly the algorithm can find a near-optimal mapping when using the different objective functions, it is important to point out that these mappings were not always reasonable when compared to the geographical region.

Another aspect to consider in evaluating the objective functions is how often the algorithm produces multiple near-optimal mappings when using different objective functions. To give an indication of how often these functions produce multiple mappings, the number of near-optimal mappings found for each state using each objective function are given in Table 3.5. The count given in the table does not include mappings that are equivalent. For example, if two mappings had the same score for a given objective function and one was merely a flipped or transposed version of the other, these mappings were only counted once because they are equivalent. Contrarily, if the two mappings had the same score but neither could be flipped or transposed to match the other mapping, then these mappings were each counted because they are unique.

The distance objective function is the only function used in the search algorithm that gives a single solution for each of the ten states. This is expected because the distance objective function is a continuous measure, while the direction and adjacent objective functions are discrete. Based on the number of multiple solutions produced by the search algorithm, the distance objective function is the most user-friendly because it will produce only one solution in most cases.

Based on all of the results, it is difficult to determine the objective function and number of random starting assignments that should be used in a particular situation because the performance of these functions is very different for each state. Since the

Table 3.5: Number of unique mapping assignments produced for ten states in the US for each objective function

State	Number of Subregions	Objective Function		
		Distance	Direction	Adjacent
Rhode Island	4	1	2	1
Connecticut	8	1	1	3
New Hampshire	8	1	2	2
Arizona	8	1	1	12
Maine	16	1	2	13
Maryland	16	1	2	16
New Mexico	32	1	12	2
California	32	1	4	1
Louisiana	64	1	1	1
Ohio	64	1	1	1

distance and direction functions require fewer random starting assignments to find a near-optimal solution, and find reasonable solutions in many cases if one considers their combined results, it is recommended that a first step should be to run the search algorithm with the distance and direction objective functions and to use Table 3.4 as a guide for chosing the number of random starting assignments. It may also be useful to increase κ or to change the stopping rule to generally improve the algorithm. If this approach fails to produce a reasonable mapping, then the algorithm should be run again using the adjacent objective function. Again, Table 3.4 should be used to determine an appropriate number of random starting assignments and it may be helpful to modify the algorithm by changing κ or the stopping rule.

It seems that the adjacent objective function should perform well when the direction and distance functions fail, as seen with New Hampshire. This did not happen, however, for New Mexico and California. It is suspected that the adjacent function failed to produce a reasonable mapping for New Mexico and California either because κ or the stopping rule were not specified appropriately or because too few random starting assignments were used. There is some evidence to support this claim. The near-optimal mapping for New Mexico in Appendix C has a score of 455 and is not considered a reasonable mapping of New Mexico, but the mapping shown in Figure 3.5 is a

reasonable mapping and has a score of 457 for the adjacent objective function. Therefore, a reasonable mapping for New Mexico exists for the adjacent objective function, the simple local search algorithm just failed to find it.

Generally, the automated mapping algorithms considered produced good maps of the geographical region within the wavelet domain, but could not be relied upon to produce the most reasonable mapping. These algorithms should be used with manual supervision. Reasonable changes to the mappings produced by these algorithms should be considered, even in cases where the changes result in a slight decrease in the value of the objective function. While the results of the simple local search algorithm are not always favorable, there is promise that this algorithm can be modified to produce reasonable mappings more frequently. There are also other search algorithms that could be explored, including the genetic algorithm of Holland (1992). The genetic algorithm was designed to avoid the problem of converging to a local optimum as opposed to a global optimum, which can occur with local search algorithms. As a result, the genetic algorithm is more complex than a local search algorithm and requires the specification of more parameters, but could offer better performance in terms of finding more reasonable wavelet domain mappings.

3.3.2 Modeling the Incidence Surface

Once a mapping of the geographical region to the wavelet domain has been determined, an incidence rate surface estimate can be obtained. In the wavelet-based surveillance method, an incidence rate surface is modeled each time an observation is collected. An observation consists of incidence counts for each subregion within the region of interest over one time interval. The counts for each subregion within a single observation are assumed to be independent. The counts for a single subregion over time are also assumed to be independent and follow a Poisson process with mean $\mu_s(i) = N_s(i)C_s(i)\lambda_s(i)$, where i is the current time interval, $\lambda_s(i)$ is the incidence rate per person in subregion s for time interval i, $N_s(i)$ is the population in subregion s for time interval i, and $C_s(i)$ is the adjustment for covariates in subregion s for time interval i.

A Poisson regression model is used to obtain an estimated incidence rate surface for each observation. Two possible Poisson regression models that can be used to estimate the incidence rate surface are discussed in Sections 3.3.2.1 and 3.3.2.2. Section 3.3.2.1

addresses the use of the Poisson regression model with the canonical link function, while Section 3.3.2.2 considers the use of the Poisson regression model with the identity link function. The estimated incidence rate surface is obtained by first estimating a mean incidence count surface using one of the Poisson regression models, where the regressors of the model are functions from the two-dimensional Haar wavelet basis. These regressors partition the wavelet domain allowing for the estimation and comparison of incidence rates over different areas of the geographical region. An example of an incidence rate surface estimate obtained from using one of the Poisson regression models, using the randomly generated observation in Figure 3.8 plot (a), is shown in Figure 3.8 plot (b). The estimated surface in Figure 3.8 plot (b) shows what an incidence surface estimate might look like for the female respiratory cancer incidence data obtained in New Mexico. The relative advantages and disadvantages of using the canonical link model versus the identity link model for estimating the incidence rate surface are discussed in Section 3.3.2.3.

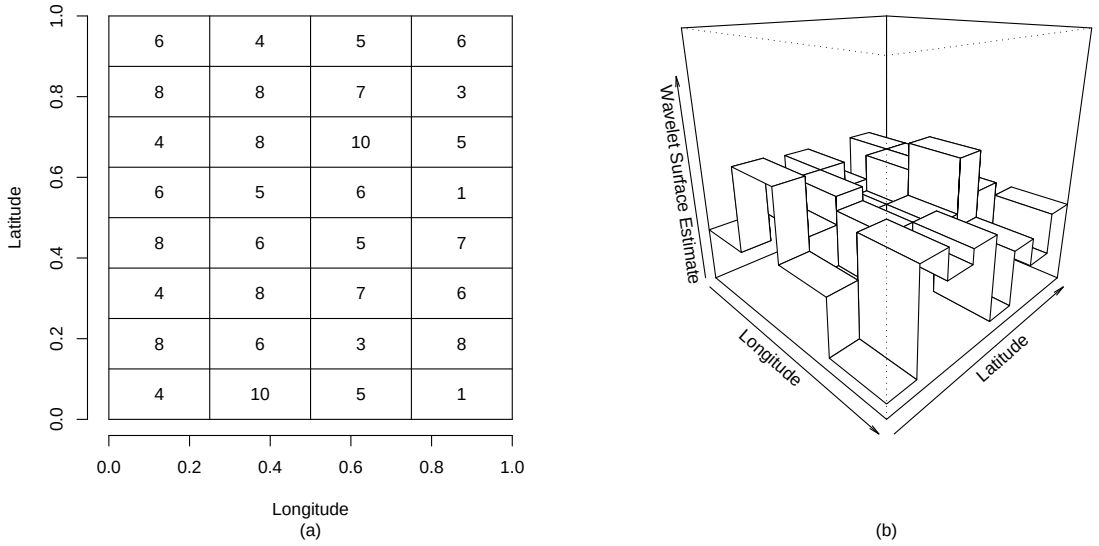


Figure 3.8: Incidence rate surface estimate example: (a) Randomly generated disease incidence counts; (b) Haar wavelet incidence surface estimate

3.3.2.1 Model using the Canonical Link Function

The form of the Poisson regression model used for estimating the mean incidence count surface when the canonical link function is used is

$$\boldsymbol{\mu}(i) = e^{ln\,[\boldsymbol{N}(i)] + ln\,[\boldsymbol{C}(i)] + \boldsymbol{X}\boldsymbol{\beta}}, \tag{3.19}$$

where i is the current time interval, $ln\,[\boldsymbol{N}(i)]$ is the $N_{wd} \times 1$ vector of the natural log of the population counts for each subregion, $ln\,[\boldsymbol{C}(i)]$ is the $N_{wd} \times 1$ vector of the natural log of the covariate adjustments for each subregion, $\boldsymbol{X} = [\mathbf{1}|\boldsymbol{\Psi}]$ is the $N_{wd} \times N_{wd}$ matrix of Haar father and mother wavelet functions, and $\boldsymbol{\beta}$ is the $N_{wd} \times 1$ vector of wavelet coefficients. The incidence rate surface estimate obtained from this model is $\hat{\boldsymbol{\lambda}}(i) = e^{\boldsymbol{X}\hat{\boldsymbol{\beta}}(i)}$, where $\hat{\boldsymbol{\lambda}}(i)$ is the $N_{wd} \times 1$ vector of estimated incidence rates in each subregion for time interval i and $\hat{\boldsymbol{\beta}}(i)$ is the $N_{wd} \times 1$ vector of estimated wavelet coefficients.

This model accounts for the effect of population and covariates on the mean incidence counts directly by including an offset for population and an offset for the covariates that affect the incidence rate of particular subregions. Covariate adjustments can be included for any factor that affects the incidence rate in the subregions, other than population. These factors could include age and gender of the residents of the subregions or environmental factors of the subregions.

Adjusting for population and covariates within the subregions by incorporating offsets into this Poisson regression model for each time interval has advantages in monitoring applications. The use of these offsets in the model is optional, but the effectiveness of the wavelet-based surveillance method to accurately detect disease clusters when these offsets are omitted may be compromised. If the population offset is not used, the counts for each observation are assumed to come from a Poisson process with mean $\mu_s(i) = C_s(i)\lambda_s(i)$, where $\lambda_s(i)$ would now be the mean incidence count in subregion s for the current time interval i. In this case an incidence rate surface can not be determined, and the mean incidence count surface would be monitored instead. This is only appropriate if the population within each subregion remains the same over the

monitoring period. Otherwise, it would be impossible to tell if a change in the mean incidence count surface is due to a shift in the population or due to the formation of a disease cluster. In most applications this assumption on population is unrealistic and information on population must be incorporated. If the covariate offset is not used, the same problem can arise. If the there are changes in the covariates within the subregions over time when this offset is excluded, a change in the monitored surface could be due to a change in the covariates or the development of a cluster.

3.3.2.2 Model using the Identity Link Function

The Poisson regression model for estimating the mean incidence count surface when the identity link function is used is

$$\boldsymbol{\mu}(i) = \boldsymbol{X}\boldsymbol{\beta}, \tag{3.20}$$

where i is the current time interval, $\boldsymbol{X} = [\mathbf{1}|\boldsymbol{\Psi}]$ is the $N_{wd} \times N_{wd}$ matrix of Haar father and mother wavelet functions, and $\boldsymbol{\beta}$ is the $N_{wd} \times 1$ vector of wavelet coefficients. The incidence rate surface for this model is found using the equation

$$\hat{\boldsymbol{\lambda}}(i) = \left(diag\left[\boldsymbol{N}(i)\right]\right)^{-1} \left(diag\left[\boldsymbol{C}(i)\right]\right)^{-1} \boldsymbol{X}\hat{\boldsymbol{\beta}}(i), \tag{3.21}$$

where $\hat{\boldsymbol{\lambda}}(i)$ is the $N_{wd} \times 1$ vector of estimated incidence rates in each subregion for time interval i, $\boldsymbol{N}(i)$ is the $N_{wd} \times 1$ vector of population counts for each subregion, $\boldsymbol{C}(i)$ is the $N_{wd} \times 1$ vector of covariate adjustments for each subregion, $diag[.]$ is an $N_{wd} \times N_{wd}$ diagonal matrix composed of the elements indicated in brackets, and $\hat{\boldsymbol{\beta}}(i)$ is the $N_{wd} \times 1$ vector of estimated wavelet coefficients. Note that the effect of population and covariates are not accounted for directly in this Poisson regression model. Instead, there is an adjustment made for these effects to obtain an incidence rate surface after the model is fit. This contrasts how the population and covariate effects are handled

using the Poisson regression model with the canonical link function, which is explained in Section 3.3.2.1. When using the identity link model, adjustments for population and covariates effects must be accounted for in the monitored statistics plotted on the control charts. If there is no adjustment made, the effects of population and the covariates would be confounded with the disease incidence rate. Therefore, if a disease cluster is detected, it would be impossible to tell whether this is due to a true cluster or changes in the population or covariates over time.

3.3.2.3 Choosing a Model

In order to select a model for use in the wavelet-based surveillance method, the relative advantages and disadvantages of the two Poisson regression models presented in Sections 3.3.2.1 and 3.3.2.2 should be considered. The main advantage of using the Poisson regression model with the identity link function is based on the average run length (ARL) performance of the control charts when this model is used as opposed to the canonical link model. The ARL performance of the control charts using both Poisson regression models is discussed in detail in Section 4.2. When monitoring statistics developed from the identity link model using the MEWMA and Weighted χ^2 control charts of Section 3.3.3, the control charts are able to detect disease clusters more quickly than the control charts using statistics derived from the canonical link model in cases where the baseline incidence counts are low. Also, the control limits determined by using the assumed approximate distributions for the MEWMA and Weighted χ^2 control charts will result in an in-control ARL closer to the nominal value when the incidence counts are low. Since the baseline incidence rates are low for many disease surveillance applications, which leads to low baseline incidence counts, the use of this model is more helpful for detecting disease clusters in many cases. The disadvantage of the identity link model, when compared to the canonical link model, is that there is no adjustment for population or covariate information in the model. This means that the values of $\hat{\boldsymbol{\beta}}(i)$ correspond to the incidence count surface and not the incidence rate surface. To account for changes in population and covariate information over time, this information must be incorporated into the statistics that are monitored in the control charts by adjusting the baseline parameter values, $\boldsymbol{\beta}_0$, for each time interval i. This means that new baseline parameter values must be determined for each time interval, which makes the method more computationally intensive.

The advantage of the Poisson regression model with the canonical link function is that it incorporates population and covariate information directly. As a result, the values of $\hat{\boldsymbol{\beta}}(i)$ correspond to the incidence rate surface and not the incidence count surface. Therefore, the baseline parameter values $\boldsymbol{\beta}_0$ remain constant over time. The use of the canonical link model with the chi-square control chart also leads to marginally better out-of-control ARL performance than the use of this chart with the identity link model, and the in-control ARLs are generally closer to the nominal value when compared to the chi-square control chart using the identity link model, except when the incidence rates are extremely low.

When selecting a model it is important to consider the incidence counts expected at baseline. If the expected incidence counts are low, then the Poisson regression model using the identity link function is generally the best choice. If the expected incidence counts are approximately 50 per 100,000 residents or larger, the ARL performance for the canonical and identity link models is similar for many monitoring scenarios. In this case, the canonical link model may be a better choice because the baseline parameter values remain constant.

3.3.2.4 Similarity to Other Methods

For a clear understanding of these models, it is important to point out how the use of wavelets in this application parallel their use for surface estimation discussed in Section 3.2. The matrix $\boldsymbol{\Psi}$, which is the matrix of regressors in the model, consists of a set of Haar mother wavelet functions for the first dimension, which represents longitude, a set of Haar mother wavelet functions for the second dimension, which represents latitude, and the crossproducts of the Haar mother wavelet functions for longitude and latitude. The expression $\boldsymbol{X}\boldsymbol{\beta} = [\mathbf{1}|\boldsymbol{\Psi}]\,\boldsymbol{\beta}$ in the model is equivalent to the function in equation (3.11) for approximating a surface using Haar wavelet functions and the use of this approximation in the Poisson regression model is analogous to its use in the regression example presented in Figure 3.4. The mother wavelet functions in the matrix $\boldsymbol{\Psi}$ are calculated by first assigning the coordinates (x_1, x_2) to the response for each subregion. The values of x_1 and x_2 are the coordinates of the center of the grid cell assigned to a subregion for the first and second dimension, respectively. These coordinates are then used to evaluate the expressions shown in equations (3.8), for $j_1 = 0, 1, \ldots, J_1$, $k_1 = 0, 1, \ldots, 2^{j_1} - 1$, $j_2 = 0, 1, \ldots, J_2$, and $k_2 = 0, 1, \ldots, 2^{j_2} - 1$.

It is also important to understand how the use of the Haar mother wavelets in this application relate to the use of dummy variables in a regression model. The purpose of using dummy variables in regression is to indicate the group of a categorical variable that a particular observation belongs to. In the wavelet-based surveillance method, the Haar wavelets are used for this same purpose. These wavelets indicate the subregion of the geographical region corresponding to a particular observation. A Poisson regression model can be fit using a set of dummy regression variables that will give the same estimated counts within each subregion as the regression model using Haar wavelets. The advantage of using the wavelet functions is that these functions indicate the location of each subregion in relation to other subregions and allow the geographical region to be subdivided at different resolutions. These multiresolution partitions produce a hierarchical breakdown of the wavelet domain into smaller and smaller areas, where the incidence rates can be modeled and changes in the incidence rates can be detected. This hierarchy can be used to improve the efficiency of a control chart to detect clusters of disease when there is a large number of model coefficients to monitor, which is discussed in Section 3.3.3.3. This hierarchy is also helpful in the implementation of diagnostic tools to determine the size and location of disease clusters once the control chart signals, which is discussed in Section 3.3.4.

3.3.3 Monitoring the Incidence Surface

To detect clusters of disease, a control chart must be selected to monitor the coefficients of the chosen Poisson regression model in equation (3.19) or (3.20). Several choices for this control chart will be discussed, which include a multivariate chi-square chart, a MEWMA chart, and a CUSUM control chart that uses a weighted χ^2 statistic. Once selected, the control chart is used to detect any change in the incidence rate surface from a baseline rate surface over time. When the control chart signals, it indicates that the incidence rate in at least one subregion in the geographical region has changed.

3.3.3.1 Multivariate Chi-Square Control Chart

First consider the multivariate chi-square control chart for monitoring the model coefficients. The purpose of a chi-square control chart is to detect mean shifts in several

variables simultaneously. A multivariate control chart is used as opposed to multiple univariate control charts, so that the correlation between the variables being monitored can be taken into account. The typical use of a chi-square control chart is in SPC applications, where q quality characteristics are monitored that follow a multivariate normal distribution with $q \times 1$ target mean vector $\boldsymbol{\mu}_0$ and $q \times q$ covariance matrix $\boldsymbol{\Sigma}$ when the process is in control. In this context, the value of a chi-square statistic is plotted on the control chart for each observation i. The chi-square statistic is

$$\chi^2(i) = n\left(\overline{\boldsymbol{x}}(i) - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1}\left(\overline{\boldsymbol{x}}(i) - \boldsymbol{\mu}_0\right), \tag{3.22}$$

where n is the sample size each time a sample is taken and $\overline{\boldsymbol{x}}(i)$ is the vector of mean characteristic estimates for observation i. This chart signals indicating a change in the mean value of at least one characteristic when $\chi^2(i) \geq \chi^2_{q,\alpha}$, where α is selected to achieve the desired in-control ARL performance. The in-control ARL is the expected number of points plotted on the control chart to obtain a signal when there has been no change in the parameter values. A signal in this case would be a false alarm.

In the wavelet-based method, a chi-square control chart is used to monitor the N_{wd} wavelet coefficients from the chosen mean incidence count model in either equation (3.19) or (3.20), as opposed to q quality characteristics. In this case, the goal is to detect a shift in one or more of the coefficients in the selected model from their baseline values, which will indicate a change in the incidence rate within some area of the geographical region. For the Poisson regression model using the canonical link, this is done by calculating the Wald statistic,

$$W(i) \;=\; \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0\right]^T \left(\boldsymbol{X}^T \hat{\boldsymbol{F}}(i) \boldsymbol{X}\right) \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0\right], \tag{3.23}$$

for each observation i, where $\hat{\boldsymbol{\beta}}(i)$ is the estimated wavelet coefficient vector for observation i, $\boldsymbol{\beta}_0$ is the vector of baseline wavelet coefficient values, and $\hat{\boldsymbol{F}}(i)$ is a $N_{wd} \times N_{wd}$

diagonal matrix with diagonal elements equal to $e^{ln[\boldsymbol{N}(i)]+ln[\boldsymbol{C}(i)]+\boldsymbol{X}\hat{\boldsymbol{\beta}}(i)}$. The value of $W(i)$ is plotted on the chi-square control chart after each observation is obtained. The statistic $W(i)$ is assumed to be approximately χ^2 with degrees of freedom equal to N_{wd} under asymptotic theory. When this assumption is made, the chart signals when $W(i) \geq \chi^2_{N_{wd},\alpha}$, where α is chosen so that the the desired in-control ARL is achieved. If this assumption is not made, then the chart signals when $W(i) \geq CL$, where CL is the value of the control limit that gives the desired in-control ARL based on the exact distribution of $W(i)$.

When the Poisson regression model with the identity link function is used, the Wald statistic plotted on the control chart is

$$W(i) = \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right]^T \left(\boldsymbol{X}^T \boldsymbol{\Sigma}_0(i)^{-1} \boldsymbol{X}\right) \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right], \tag{3.24}$$

where $\boldsymbol{\beta}_0(i)$ is the vector of baseline wavelet coefficient values for observation i, $\boldsymbol{\Sigma}_0(i)$ is the $N_{wd} \times N_{wd}$ covariance matrix of the baseline incidence counts for observation i, and all other variables have the same definitions given for the canonical link model. Notice that the baseline parameter vector, $\boldsymbol{\beta}_0(i)$, is dependent on the observation i. This is due to the fact that the baseline parameter vector must change for each observation based on the changes in the population and covariates as discussed in Section 3.3.2.2. When the Wald statistic based on the identity link model is used, the control chart signals under the same circumstances as when the Wald statistic based on the canonical link model is plotted.

In the wavelet-based method, the chi-square control chart is used prospectively, but this chart can be used in both retrospective and prospective analyses. In retrospective analyses, chi-square control charts are typically used to evaluate historical data in order to determine baseline parameter values. In prospective analyses, chi-square charts are used to detect shifts in the monitored parameters from the baseline or target values. One disadvantage of using a chi-square control chart in a prospective analysis is that its strength is in detecting large shifts in the parameter values. Therefore it is important to consider the use of the MEWMA control chart of Lowry *et al.* (1992), which is better for detecting small shifts in the parameters.

3.3.3.2 MEWMA Control Chart

In many applications, including the detection of disease clusters, it is important to detect small shifts from baseline indicating an out-of-control state before a larger shift occurs. The MEWMA control chart of Lowry *et al.* (1992) has proven to be more efficient in detecting small shifts in a parameter vector. Therefore, an MEWMA control chart, that has been adjusted to monitor coefficients from the selected Poisson regression model, is suggested when there is a need to detect small shifts from the baseline parameter vector.

The statistic plotted on the MEWMA control chart for time interval i when the canonical link Poisson regression model is used is

$$MEWMA(i) \;=\; \boldsymbol{z}(i)^T \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{z}(i)} \boldsymbol{z}(i), \tag{3.25}$$

where

$$\boldsymbol{z}(i) \;=\; \lambda \left(\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0 \right) + (1-\lambda)\, \boldsymbol{z}(i-1), \tag{3.26}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}(i)} \;=\; \frac{\lambda}{2-\lambda} \left(\boldsymbol{X}^T \hat{\boldsymbol{F}}(i) \boldsymbol{X} \right)^{-1}, \tag{3.27}$$

and λ is the MEWMA control chart parameter. The value of λ must be chosen so that $0 \leq \lambda \leq 1$. A standard choice for λ is 0.2 and this value will be used in the forthcoming demonstration and evaluation of this control chart as part of the the wavelet-based surveillance method. This control chart signals indicating the incidence rate surface has changed when $MEWMA(i) \geq CL$. The value of CL can be selected to achieve an approximate desired in-control ARL by assuming that the parameter vector follows a multivariate normal distribution, with mean vector $\boldsymbol{\beta}_0$ and covariance matrix $\left(\hat{\boldsymbol{F}}(i) \boldsymbol{X} \right)^{-1}$, when the incidence rate is in-control, or by determining the value

that would give the exact desired in-control ARL based on simulation. Values for CL assuming multivariate normality are readily available in statistical computer packages, which would allow immediate execution of a control chart. To obtain CL values for an exact in-control ARL, a simulation would have to be designed and run prior to implementing a control chart.

When the identity link Poisson regression model is used, the statistic plotted on the MEWMA control chart for time interval i is

$$MEWMA(i) = \boldsymbol{z}(i)^T \boldsymbol{\Sigma}_{\boldsymbol{z}(i)}^{-1} \boldsymbol{z}(i), \tag{3.28}$$

where

$$\boldsymbol{z}(i) = \lambda \left(\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i) \right) + (1 - \lambda)\, \boldsymbol{z}(i-1) \tag{3.29}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{z}(i)} = \frac{\lambda}{2 - \lambda} \left(\boldsymbol{X}^T \boldsymbol{\Sigma}_0(i)^{-1} \boldsymbol{X} \right)^{-1}. \tag{3.30}$$

The value chosen for the MEWMA parameter, λ, is also 0.2 in this case and the control chart will signal under the same conditions as the control chart using the MEWMA statistic based on the canonical link model. The control limit value, CL, can be determined both approximately and by simulation in the same way the value is determined when the canonical link model is used.

3.3.3.3 Weighted χ^2 Control Chart

Instead of using a chi-square or MEWMA control chart, it is also of interest to consider the use of a Weighted χ^2 control chart, where a weighted χ^2 statistic, developed from the Wald statistic in equation (3.23) or (3.24), is monitored. This type of chart is most advantageous in applications where there is a desire to place more importance

on the detection of clusters of a particular size. For instance, one may want to put more emphasis on the detection of large clusters over the detection of more localized clusters, or vice versa. If this is the case, a statistic should be used that has more power to detect disease clusters of the specified size. This can be accomplished by weighting the contribution of the wavelet coefficients in equation (3.23) or (3.24) based on their resolution. For the Poisson regression model using the canonical link function, this results in a weighted χ^2 statistic of the form

$$WCHISQ(i) = \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0\right]^T \boldsymbol{W}^{\frac{1}{2}} \left(\boldsymbol{X}^T \hat{\boldsymbol{F}}(i) \boldsymbol{X}\right) \boldsymbol{W}^{\frac{1}{2}} \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0\right], \quad (3.31)$$

where $\boldsymbol{W}$ is a diagonal matrix with the weights for the wavelet coefficients on the diagonal and i is the current time interval. For the Poisson regression model using the identity link function, the weighted χ^2 statistic is

$$WCHISQ(i) = \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right]^T \boldsymbol{W}^{\frac{1}{2}} \left(\boldsymbol{X}^T \boldsymbol{\Sigma}_0(i)^{-1} \boldsymbol{X}\right) \boldsymbol{W}^{\frac{1}{2}} \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right] \quad (3.32)$$

where $\boldsymbol{W}$ is again a diagonal matrix with weights corresponding to the wavelet coefficients on the diagonal and i is the current time interval. When using these statistics, the wavelet coefficients with resolution corresponding to the cluster size of importance should be weighted more heavily.

An approximate distribution must be determined for the weighted statistics in equations (3.31) and (3.32) before they can be used in control charts for prospective monitoring. Even though these statistics are modifications of the Wald statistics in equations (3.23) and (3.24), respectively, they do not have an approximate chi-square distribution with N_{wd} degrees of freedom at baseline because they are weighted quadratic forms. The transformation of Wilson and Hilferty (1931) can be used to transform random variables from the distribution of a definite quadratic form into random varibles that follow an approximate standard normal distribution. Using this transformation to ap-

proximately normalize the weighted χ^2 statistics is useful in this application because existing ARL results, based on the standard normal distribution, can then be used for control chart design.

A definite quadratic form is one that can be expressed as $Q_k(\boldsymbol{c}, \boldsymbol{a}) = \sum_{j=1}^{k} c_j (x_j + a_j)^2$, where $\boldsymbol{x}$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix I and $\boldsymbol{c}$ and $\boldsymbol{a}$ are vectors of constants such that $c_j > 1$ and $1 \leq j \leq k$. If $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ is the spectral decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\frac{1}{2}} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the covariance matrix of $\boldsymbol{\beta}$, then the quadratic forms in equations (3.31) and (3.32) can be written as the definite quadratic form

$$Q_{N_{wd}}(i) = \sum_{j=1}^{N_{wd}} \lambda_j \hat{\gamma}_j \tag{3.33}$$

where λ_j is the j^{th} diagonal element of the matrix $\boldsymbol{\Lambda}$ and $\hat{\gamma}_j$ is the j^{th} element of the vector $\hat{\boldsymbol{\gamma}} = \boldsymbol{U}^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\frac{1}{2}} \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0\right]$ when the canonical link model is used or the j^{th} element of the vector $\hat{\boldsymbol{\gamma}} = \boldsymbol{U}^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\frac{1}{2}} \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right]$ when the identity link model is used. The vector $\hat{\boldsymbol{\gamma}}$ has an approximate multivariate normal distribution with mean vector $\mathbf{0}_{N_{wd} \times 1}$ and covariance matrix $I_{N_{wd}}$ since $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\frac{1}{2}} \left[\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}_0(i)\right]$ is approximately multivariate normal with mean vector $\mathbf{0}_{N_{wd} \times 1}$ and covariance matrix $I_{N_{wd}}$ and $\boldsymbol{U}^T\boldsymbol{U} = I_{N_{wd}}$. Therefore, $Q_{N_{wd}}(i)$ can be transformed into a standard normal random variable using the Wilson-Hilferty transformation

$$z(i) = \frac{\left[\frac{Q_{N_{wd}}(i)}{\theta_1}\right]^{h_0} - 1 - \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2}}{\frac{h_0 \sqrt{2\theta_2}}{\theta_1}}, \tag{3.34}$$

where

$$\theta_s = \sum_{j=1}^{N_{wd}} \lambda_i^s \tag{3.35}$$

and

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}. \tag{3.36}$$

Once the weighted χ^2 statistic has been transformed it can be monitored using a CUSUM chart designed for monitoring standard normal random variables. The use of CUSUM charts is discussed in detail in Section 2.2.2. In the wavelet-based monitoring method, the values plotted on the CUSUM chart are

$$C_j = max\left(0, z(j) - \frac{1}{2} + C_{j-1}\right), \tag{3.37}$$

for $j = 1, 2, \ldots, i$, where i is the current observation and $C_0 = 0$. This control chart will be referred to as the Weighted χ^2 control chart. This chart will signal, indicating the incidence rate surface has changed, when $C_j \geq h$. The value of h can be chosen to achieve an approximate specified in-control ARL by assuming that $z(j)$ follows a standard normal distribution or to achieve an exact specified in-control ARL by simulation, as with the MEWMA control chart discussed in Section 3.3.3.2.

While the Weighted χ^2 control chart is effective for detecting clusters of a specified size, care must be taken when choosing to use this control chart in monitoring applications. If a cluster develops that is a different size than expected, and the weights are assigned to detect clusters of the expected size, the chart will have less power to detect the cluster and may miss it altogether. Therefore, this chart is recommended for use in cases when detecting clusters of a particular size is of interest.

Another application of the Weighted χ^2 control chart one can consider is to attempt to choose the weights of the monitored statistic so that this control chart is more efficient in detecting changes in the incidence rate surface than either a chi-square or an MEWMA control chart. When using either the chi-square or MEWMA chart, it becomes more difficult to detect shifts in the parameter vector as the number of pa-

rameters being monitored increases. By using either of the weighted χ^2 statistics in equations (3.31) and (3.32), certain dimensions of the parameter space can be emphasized, which can make it easier to detect shifts in parameters that are considered the most important. It seems reasonable that the Weighted χ^2 control chart could be used to improve the efficiency of cluster detection in the wavelet-based method by weighting the wavelet coefficients that partition larger sections of the wavelet domain more heavily. These are the lower resolution coefficients. This seems like a reasonable approach because, in many cases, changes in small partitions of the wavelet domain will impact the values of the low resolution coefficients, as well as the high resolution coefficients. A simple weighting scheme for the weighted χ^2 statistics in this application is to weight each coefficient by taking the inverse square-root of its resolution increased by one. This scheme has been shown to perform well in the class of tests based on weighted quadratic forms by Spitzner (2008), where these weights are applied to individual coefficients. This application is slightly different, however, since the weights are being applied to groups of coefficients of the same resolution instead of individual parameters. When demonstrating the Weighted χ^2 control chart in Section 3.4 and evaluating this control chart in Section 4.2, this application will be the focus rather than the use of this control chart to detect clusters of a specific size. While this strategy seems advantageous, the evaluation of this control chart will show that this method does not perform as well as expected. This could be a result of the weighting scheme used because the performance of the weighting scheme was unknown for this type of application. The performance of this method will be discussed in more detail in Section 4.2.

3.3.4 Determining Cluster Size and Location

Once the control chart signals, diagnostics can be used to determine the size and shape of the area or areas in the geographical region where there has been a change in the incidence rates. These diagnostics accomplish this by determining the maximum resolutions needed in both dimensions of the incidence rate surface to describe the change in the incidence rates. These maximum resolution values aid in indicating the size and shape of a possible cluster. The diagnostics can also be used to help determine

the location of a possible disease cluster, because they determine the mother wavelet functions needed in the mean incidence model to adequately describe the change in the incidence rates. Once the necessary wavelet functions are determined, the model including only the corresponding wavelet coefficients can be used to estimate the change in the incidence rate in each subregion from the baseline value. These estimates can then be used to create a shaded map of the geographical region indicating the intensity of the estimated change in the incidence rates from the baseline rates.

Two diagnostics are suggested for the purpose of determining the wavelet coefficients needed in the mean incidence model to describe changes in the incidence rate surface. These are the standardized Wald statistic (SWS) and Akaike's information criterion (AIC). In a typical regression setting, these diagnostics can be used to determine the parameters in a regression model that are important for describing the relationship between the response and explanatory variables, and that will be useful for obtaining predictions. In the wavelet-based method, either of these diagnostics can be used to determine the coefficients in the mean incidence surface model that are important for estimating the changes in the incidence rates of the subregions. This is done by calculating the SWS or AIC for each possible reduced resolution model and the full resolution model, so that these models can be compared. The SWS is used to standardize the weighted χ^2 statistic so that it has mean zero and variance equal to one for each reduced resolution model. This diagnostic is based on the Adaptive Neyman Test of Fan (1996), which is a commonly used test statistic in functional data analysis. The SWS for the reduced resolution models in this application is

$$SWS = \frac{W_R(i) - p_R}{\sqrt{2p_R}}, \tag{3.38}$$

where p_R is the number of wavelet coefficients in the reduced model and

$$W_R(i) = \left[\hat{\boldsymbol{\beta}}_R(i) - \boldsymbol{\beta}_{BR}\right]^T \left(\boldsymbol{X}_R^T \hat{\boldsymbol{F}}_R(i) \boldsymbol{X}_R\right) \left[\hat{\boldsymbol{\beta}}_R(i) - \boldsymbol{\beta}_{BR}\right], \tag{3.39}$$

for the Poisson regression model using the canonical link function. In equation (3.39), $\hat{\boldsymbol{\beta}}_R(i)$ is the vector of p_R estimated wavelet coefficients for the reduced model for time interval i, $\boldsymbol{\beta}_{BR}$ is the vector of p_R baseline coefficient values for the reduced model, $\boldsymbol{X}_R = [\mathbf{1}|\boldsymbol{\Psi}_R]$, and $\hat{\boldsymbol{F}}_R(i)$ is a $p_R \times p_R$ diagonal matrix with diagonal elements equal to $e^{ln[\boldsymbol{N}(i)]+ln[\boldsymbol{C}(i)]+\boldsymbol{X}_R\hat{\boldsymbol{\beta}}_R(i)}$. For the Poisson regression model using the identity link function,

$$W_R(i) \;=\; \left[\hat{\boldsymbol{\beta}}_R(i) - \boldsymbol{\beta}_{BR}(i)\right]^T \left(\boldsymbol{X}_R^T \boldsymbol{\Sigma}_{BR}^{-1}(i) \boldsymbol{X}_R\right) \left[\hat{\boldsymbol{\beta}}_R(i) - \boldsymbol{\beta}_{BR}(i)\right], \qquad (3.40)$$

where $\boldsymbol{\beta}_{BR}(i)$ is the vector of p_R baseline coefficient values for the reduced model for time interval i, $\boldsymbol{\Sigma}_{BR}^{-1}(i)$ is the $p_R \times p_R$ covariance matrix of the baseline incidence counts for the reduced model for observation i, and the other variables are defined as in equation (3.39).

Instead of standardizing the weighted χ^2 statistic of each reduced model, the AIC diagnostic applies a penalty to the weighted chi-square statistic so that models with a larger number of wavelet coefficients are penalized more for their complexity. This idea is based on the standard regression AIC diagnostic introduced by Akaike (1974). The AIC for the reduced resolution models in this application is

$$AIC \;=\; W_R(i) - 2p_R, \qquad (3.41)$$

where p_R and $W_R(i)$ are defined as in equations (3.38), (3.39), and (3.40). The SWS and AIC are calculated for the full model by replacing $W_R(i)$ with $W(i)$, from equation (3.23) or (3.40), and by replacing p_R with N_{wd} in equations (3.38) and (3.41), respectively. The SWS and AIC diagnostics evaluate the adequacy of a particular model to give accurate estimates of the changes in the incidence rates while accounting for the number of coefficients in the model. A model that has a large value of either the SWS or AIC is one that will be good for describing changes in the incidence rates. In the

wavelet-based method, the model that is used to estimate changes in the incidence rates will be the model that gives the maximum value for either the SWS or AIC. In addition to determining the model with the maximum SWS or AIC, it may also be of interest to determine other models that produce large values of these diagnostics. A plot of the diagnostic values for these models, grouped by their dimension resolution combination, can be used for this purpose. A plot of this type should also confirm that the resolution combination of the selected model is appropriate, because one would expect high values of the diagnostics for all models with this same resolution combination.

The reduced resolution models compared using either the SWS or AIC are models that are similar to the full model used for estimating the incidence rate surface, but they do not contain all of the wavelet coefficients. These models contain a subset of the wavelet coefficients that do not partition the wavelet domain down to the subregion level as the full model does. The reduced resolution models produce lower resolution partitions of the wavelet domain, which leave some subregions grouped together as a whole. To illustrate how these reduced models partition the wavelet domain, 39 examples of these partitions are given in Figure 3.9 for a geographical region containing 32 subregions, such as New Mexico. In the reduced model partitions shown in Figure 3.9, the incidence rate estimate for all subregions in each cell created by partitioning the wavelet domain is equivalent. The full model partition is also given in Figure 3.9 in the lower right-hand corner. The models for each of the 40 partitions shown in Figure 3.9 will allow for changes in the incidence rates in each separated area of the partition. There are a total of 129 reduced models and corresponding partitions for a geographical region with 32 subregions with resolutions of one and two for longitude and latitude. Each of these 129 reduced models, as well as the full model, would be compared using the SWS or AIC when selecting a model for estimating changes in the incidence rates for New Mexico, or any geographical region similar to New Mexico.

When the appropriate model has been selected using one of the diagnostics, this model can be used to obtain estimates of the changes in the incidence rates in the geographical region under surveillance. These estimates are used to shade the map of the geographical region according to the intensity and direction of the changes in the incidence rates. By examining this map, clusters of disease can be identified by determining the subregions with highly elevated incidence rates, when compared to

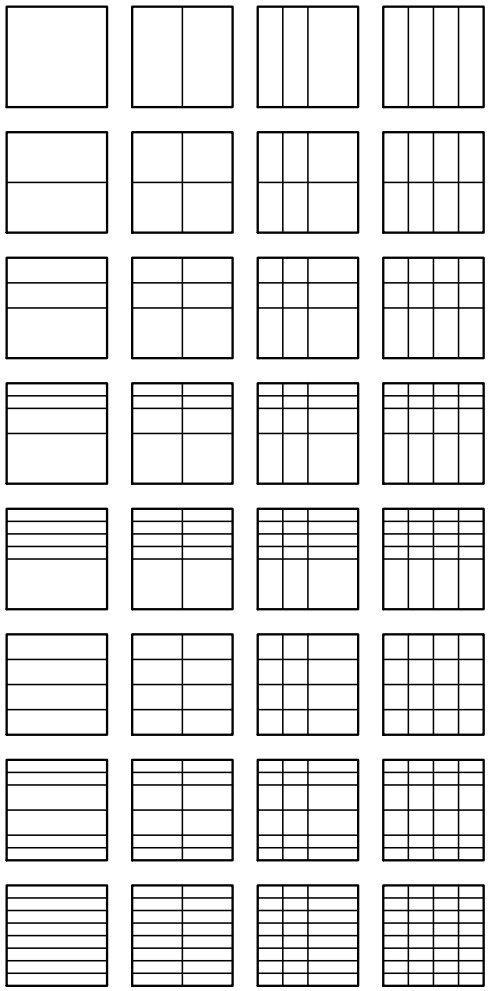


Figure 3.9: Multiresolution partitions of the two dimensional Haar wavelet domain for a geographical region containing 32 subregions

other subregions. There is also the possibility that the chart signaled due to a decrease in the incidence rates in some subregions. In this case, areas with large decreases in the incidence rates can be identified. This reduced-model approach can be more beneficial than merely using estimates from the full model to create a shaded map of the incidence rate estimates within each subregion. The reduced models help to eliminate noise present in the raw data, thereby highlighting clusters and clarifying their shapes. An example of a shaded map used to identify clusters in New Mexico is shown in Figure 3.10.

In cases where a disease cluster spans over a geographical area that is represented by several high resolution wavelet functions in the Poisson regression model, the reduced models that maximize the SWS and AIC statistics may not adequately describe the changes in the incidence rate surface. When this occurs, the reduced models selected tend to over-smooth the incidence rate surface estimate and a cluster may appear to cover more subregions than it truly does. This can make it difficult to determine the location of a cluster using the SWS and AIC diagnostics in some scenarios. An example of this can be seen in the first demonstration shown in Section 3.4. To avoid issues determining the cluster location using these diagnostics, it may be helpful to consider other procedures for determining a reduced model that estimates the incidence rate surface well. One possibility is to use stepwise regression techniques to select the wavelet coefficients that are beneficial for estimating the incidence rate surface. If the SWS or AIC diagnostics are used, however, it is helpful to obtain estimates of the changes in the incidence rates from baseline using the full model. Then these estimates can be used to produce a shaded map of the changes in incidence rates from baseline over the geographical region. Since the full model is saturated, this provides a map comparing the current incidence rates in each subregion to the baseline rates. The rates in these maps tend to vary widely from baseline in some subregions, even when there is no change in the mean incidence rate surface. Therefore, these maps should not be used alone because there is danger of falsely identifing a cluster in a geographical area that has an increased incidence rate due to random variation. It is best to use the shaded maps from the SWS and AIC reduced models in conjunction with the shaded map from the full model.

Another issue can arise when using these diagnostics that can make it difficult to determine the location of a disease cluster within a geographical region. This issue

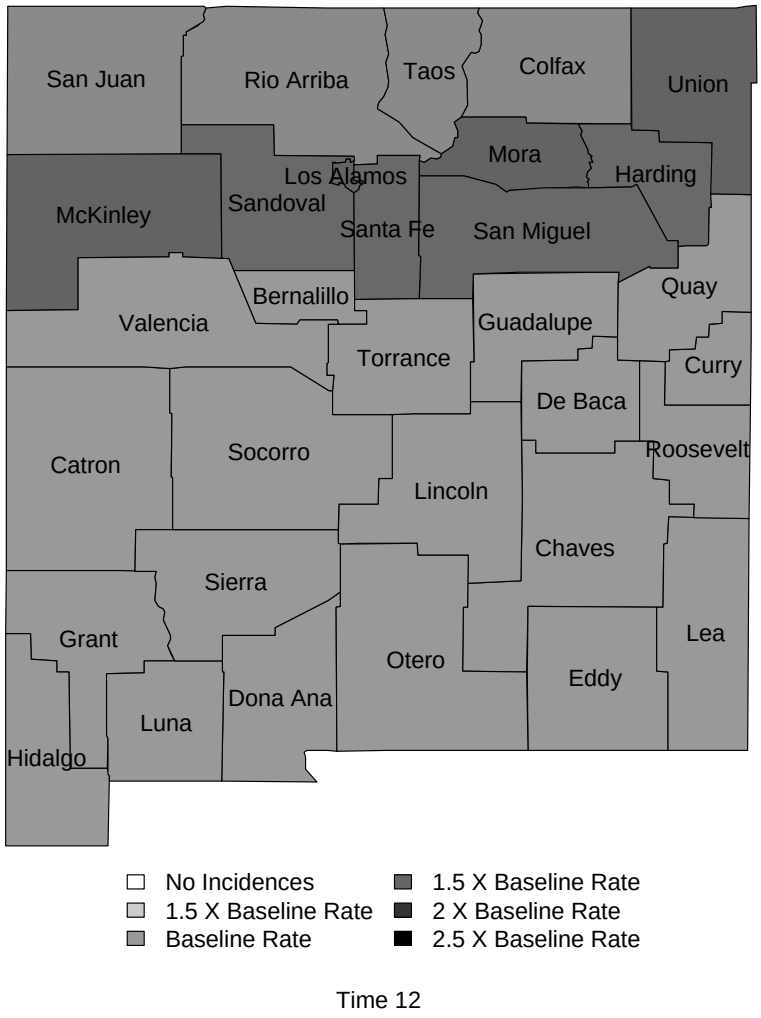


Figure 3.10: Example map of the ratios of estimated incidence rates to baseline in the counties of New Mexico

occurs when the shift in the incidence rate surface is detected by the MEWMA or Weighted χ^2 contol chart from cumulative observations that contribute to a signal as opposed to an individual independent observation that is extreme enough to cause a signal alone. In this case, the diagnostics calculated using the observation at the time of the signal may not provide a good indication of the cluster location. This happens because the subregions with mean incidence rates that differ from baseline leading up to the signal may not have incidence rates that differ substantially from baseline at the time of a signal due to natural variation. An example of this can be seen in the third demonstration shown in Section 3.4. Although a formal solution to address this issue is not developed here, it seems that a reasonable approach would be to modify the calculation of the diagnostics so that they could incorporate information from observations obtained prior to a signal. The information from past observations could be weighted as it is in the MEWMA and Weighted χ^2 contol charts. It may also be helpful to look at the raw data from the past observations to determine areas that had increased incidence rates leading up to a signal.

3.4 Demonstration of the Surveillance Method

To illustrate the wavelet-based disease surveillance method, demonstrations of the method are shown for three different simulated scenarios. There are two purposes for presenting these demonstrations. One is to give an example of how the method can be applied. The other is to show the types of clustering this method can detect.

These three demonstrations were done using simulated data for a geographical region consisting of 32 subregions, where an assumption was made that the mapping of these subregions to the wavelet domain had already been determined. Thirty-two subregions were used to represent the counties in the state of New Mexico, since the examples shown up to this point have been in the context of the female respiratory lung cancer data collected in this region. The population in each subregion was assumed to be 100,000 residents. The baseline incidence rate per person was assumed to be the same for each subregion and no covariate adjustments were used in the demonstrations. These assumptions make it easier to see where there is a subregion with an increased or decreased incidence rate, because the incidence rate surface at baseline is flat. In a

demonstration where the baseline incidence rates are different within each subregion, the results would be similar to those shown. Each demonstration was run for 20 time intervals, where the incidence rate surface changed from baseline at time interval 11.

In each demonstration the incidence rate surface was changed to illustrate a different situation when a disease cluster is present. This was done by changing the values of the wavelet coefficients. The wavelet coefficients describe the overall mean of the incidence rate surface, as well as the shape of the surface. It is important for the method to detect clusters present when there is a change in the overall mean of the surface, the shape of the surface, or both. Therefore, the changes in the coefficients of the incidence rate surface in the demonstrations correspond to these situations. In the first demonstration, the incidence rate increased in four subregions, while the incidence rate remained the same in all other subregions. This resulted in a change in both the overall mean and shape of the incidence rate surface. In the second demonstration, the incidence rate increased in all of the subregions, which only changed the overall mean of the surface. In the third demonstration, the incidence rate increased in four subregions and also decreased in four subregions such that there was no change in the overall mean of the surface. This resulted in a change in the shape of the incidence rate surface only.

In each scenario, the baseline coefficient vector, $\boldsymbol{\beta}(B)$ shifted to an out-of-control coefficient vector, denoted by $\boldsymbol{\beta}_1$. The statistics used to monitor the vector of coefficients were the Wald statistics for the canonical and identity link models in equations (3.23) and (3.24), respectively, the MEWMA statistics for the canonical and identity link models in equations (3.25) and (3.28), respectively, and the weighted χ^2 statistics for the canonical and identity link models in equations (3.31) and (3.32), respectively.

So that each demonstration is comparable, the magnitude of the change in the coefficients of the Poisson regression model using the identity link function was made the same in each case. This was done through the use of a noncentrality parameter, calculated using the baseline and out-of-control coefficient vectors. The noncentrality parameter used was

$$\delta = \frac{1}{2}\left[\boldsymbol{\beta}_1 - \boldsymbol{\beta}(B)\right]^T \boldsymbol{\Sigma}\left(\mathbf{pc}\boldsymbol{\lambda}_0\right)^{-1}\left[\boldsymbol{\beta}_1 - \boldsymbol{\beta}(B)\right], \tag{3.42}$$

where

$$\boldsymbol{\Sigma}(\mathbf{pc}\boldsymbol{\lambda}_0)^{-1} \;=\; \boldsymbol{X}^T diag\left[\mathbf{pc}\boldsymbol{\lambda}_0\right]^{-1}\boldsymbol{X}, \tag{3.43}$$

$\mathbf{pc}\boldsymbol{\lambda}_0$ is the vector of counts at baseline based on the assumed population size and covariate adjustment in each subregion, and $diag[.]$ is an $N_{wd} \times N_{wd}$ diagonal matrix composed of the elements indicated in brackets. This noncentrality parameter was chosen over a comparable noncentrality parameter related to the Poisson regression model using the canonical link function because the identity link model has better ARL performance in scenarios similar to those in these demonstrations as shown in Section 4.2. Therefore, the magnitude of the shifts in the coefficients for the Poisson regression model using the canonical link function were only appoximately the same in the demonstrations. The value of the noncentrality parameter used in the demonstrations was $\delta = 3$.

The purpose of the first demonstration is to show that the control charts presented in Section 3.3.3 can effectively detect a cluster forming in a portion of the subregions within the geographical region of interest, which affects both the overall mean and shape of the incidence rate surface. In this demonstration, the incidence rate increased in four of the 32 subregions, while the incidence rate remained the same in all other subregions. The baseline incidence rate was 10 incidences per 100,000 residents, and the mean incidence count remained at baseline in all subregions for time intervals 1 to 10. At time interval 11, a disease cluster formed increasing the incidence rate to 13.873 per 100,000 residents in four adjacent subregions in the northern part of the geographical region. The incidence rate remained at 13.873 per 100,000 residents in these subregions through time interval 20, while the other subregions maintained an incidence rate of 10 per 100,000 residents. The resulting incidence rate surfaces for these time intervals are shown in Figure 3.11. The change in the surface at time interval 11 caused the intercept to increase and the coefficients associated with the incidence rate in the four subregions to shift. The mean incidence rate over all subregions increased to 10.484 per 100,000 residents from 10 per 100,000 residents.

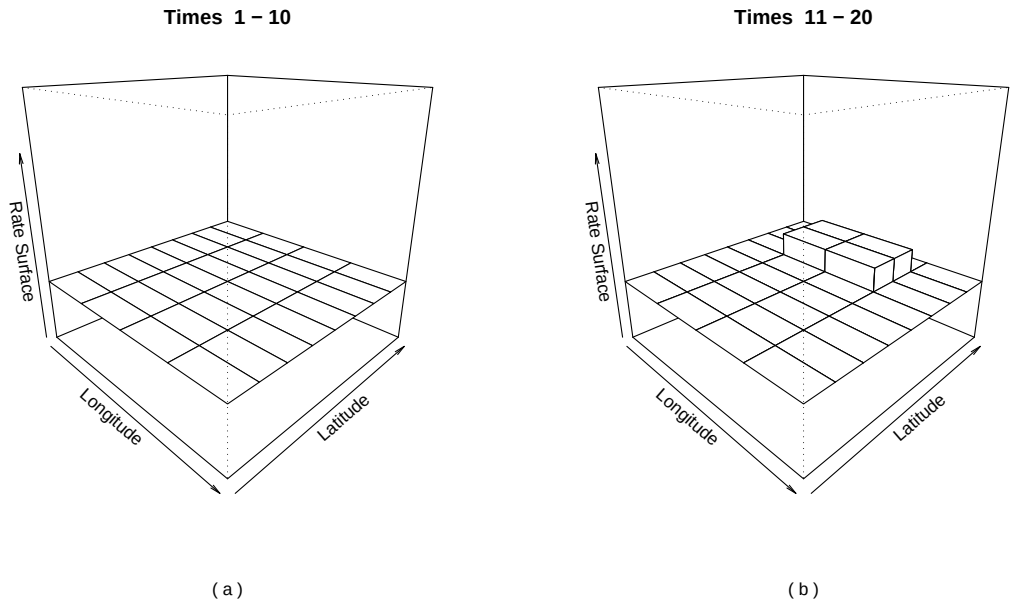


Figure 3.11: Incidence rate surfaces for the first demonstration: (a) Incidence rate surface for time intervals 1 to 10; (b) Incidence rate surface for time intervals 11 to 20

The control charts presented in Section 3.3.3 were used to prospectively monitor the region over the 20 time intervals. Random incidence count data following a Poisson distribution were generated based on the incidence rate surfaces in Figure 3.11 for each time interval. The applicable statistics for each control chart, for $i = 1, 2, \ldots, 20$, were calculated and plotted on the charts as each observation was obtained. The control limits used for these charts were 56.33, 54.85, and 3.502 for the chi-square, MEWMA, and Weighted χ^2 control charts, respectively. Each of these control limits results in an approximate in-control ARL of 200. When monitoring monthly data, such as monthly female respiratory lung cancer incidences, one would expect a false alarm approximately once every 200 months. The final chi-square, MEWMA, and Weighted χ^2 control charts using the identity link Poisson regression model for this scenario are shown in Figures 3.12, 3.13, and 3.14, respectively. The full demonstration of the method for this scenario, including the simulated observations and intermediate incidence surface estimates for each time interval, is provided in Appendix D. The control charts for the the canonical link model are also shown in Appendix D. For the

identity link model, each of the control charts signaled at time interval 12 indicating that the incidence rate surface changed from baseline. When using the canonical link model on the same simulated data, all three control charts also signaled, but at a later time point.

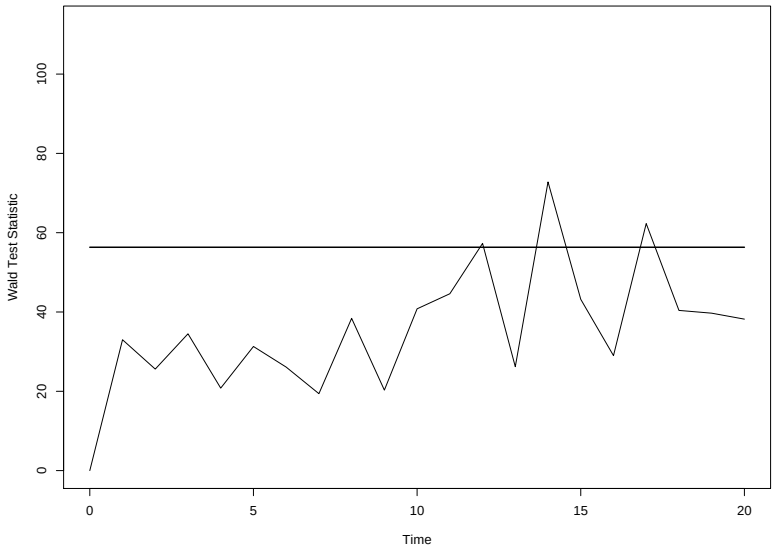


Figure 3.12: Chi-square control chart for the first demonstration

To demonstrate how the location of a cluster can be found once a control chart signals, the SWS and AIC diagnostics based on the identity link model were used to find reduced models to estimate the changes in the incidence rate surface when the first signal occurred for time interval 12. The shaded maps of the estimated changes in the incidence rates from baseline over the wavelet domain using the SWS and AIC diagnostics are shown in Figures 3.15 and 3.16, respectively. In this demonstration, the disease cluster is in a geographical area where the incidence rates can only be estimated adequately using the high resolution wavelet functions in the Poisson regression model. Therefore, the SWS and AIC reduced models are over-smoothing the incidence rate surface estimate as discussed in Section 3.3.4. The shaded map of the estimated changes in the incidence rates from baseline using the full identity link model is provided in Figure 3.17 to aid in determining the location of the disease cluster. This plot shows

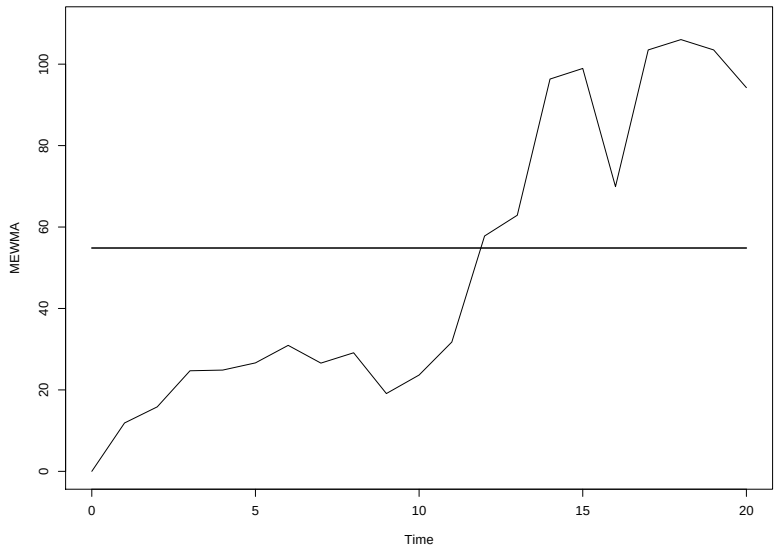


Figure 3.13: MEWMA control chart for the first demonstration

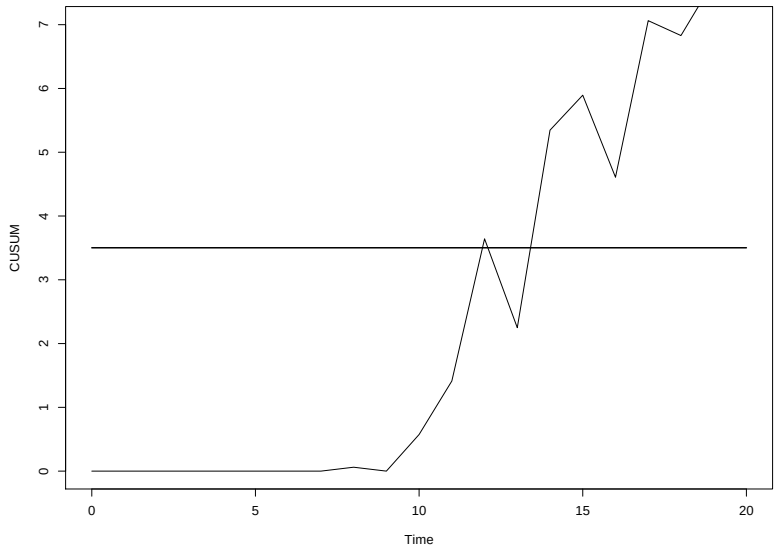


Figure 3.14: Weighted χ^2 control chart for the first demonstration

the four subregion cluster in the northern part of the region clearly, but also shows other subregions with elevated incidence rate estimates for time interval 12 because this plot shows all of the noise present in the raw data. Using all three diagnostic plots in conjunction, one might look for the subregions that have an increased incidence rate estimate in all plots. Using this approach, the four subregions in the center of the northern half of the geographical region would be correctly identified as part of the disease cluster. One might also consider the subregion with an elevated incidence rate estimate in the northernmost part of the region, just above the true cluster, part of the disease cluster.

The second demonstration is provided to show that the control charts presented in Section 3.3.3 are capable of detecting a cluster forming over the entire geographical region of interest, where there is an equivalent increase in the incidence rate in each subregion and the shape of the incidence rate surface remains the same. This demonstration was done similarly to the first demonstration. The geographical area of interest consisted of 32 subregions and the assumptions remained the same, but in this case, when the incidence rate surface changed at time interval 11, the surface shifted to one different than in the previous demonstration. In this case, the incidence rates remained at a baseline value of 10 per 100,000 residents for time intervals 1 to 10. At time interval 11, the incidence rate increased to 11.369 per 100,000 residents over the entire region of interest and the incidence rate stayed constant at the value 11.369 per 100,000 residents through the final time interval 20. The incidence rate surfaces for this case are shown in Figure 3.18. In this case, the change in the surface from baseline at time interval 11 only influenced the intercept.

This second situation shows how the control charts in Section 3.3.3 respond when there is a constant increase in the incidence rate over the whole geographical region. The control charts were designed and used in the same way in this demonstration as they were in the first demonstration. Resulting chi-square, MEWMA, and Weighted χ^2 charts using the Poisson regression model with the identiy link function are shown in Figures 3.19, 3.20, and 3.21, respectively, for randomly generated Poisson incidence counts. The MEWMA and Weighted χ^2 control charts both signaled at time interval 15, while the chi-square control chart signaled later at time interval 17. All three charts indicated that the incidence rate surface changed. Although not shown, the chi-square,

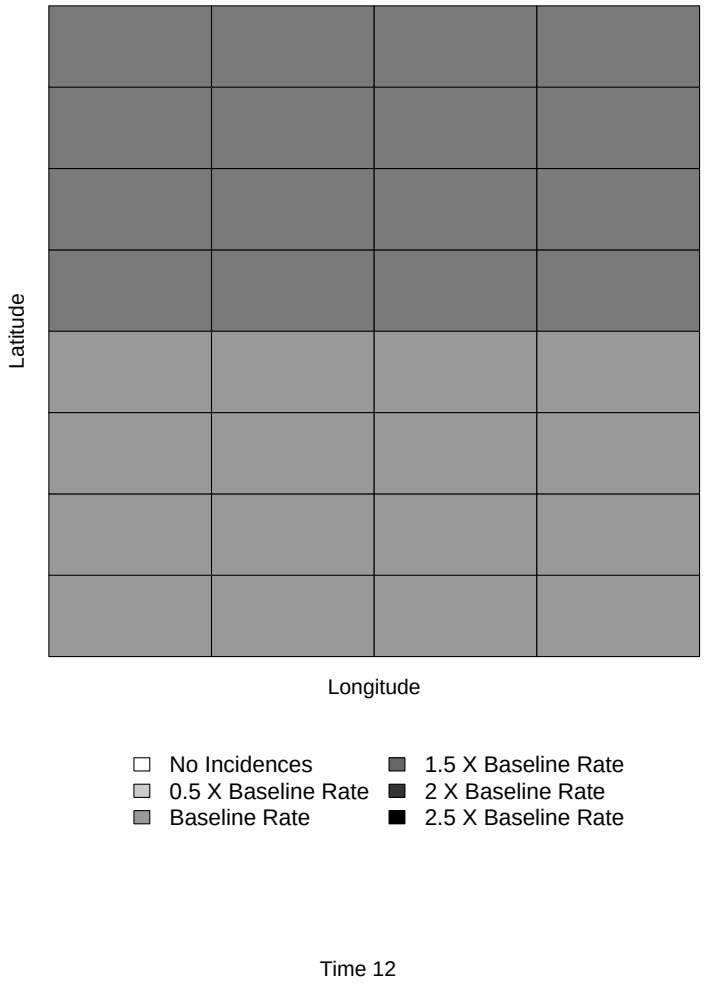


Figure 3.15: Ratios of estimated incidence rates to baseline for the first demonstration using the SWS reduced model

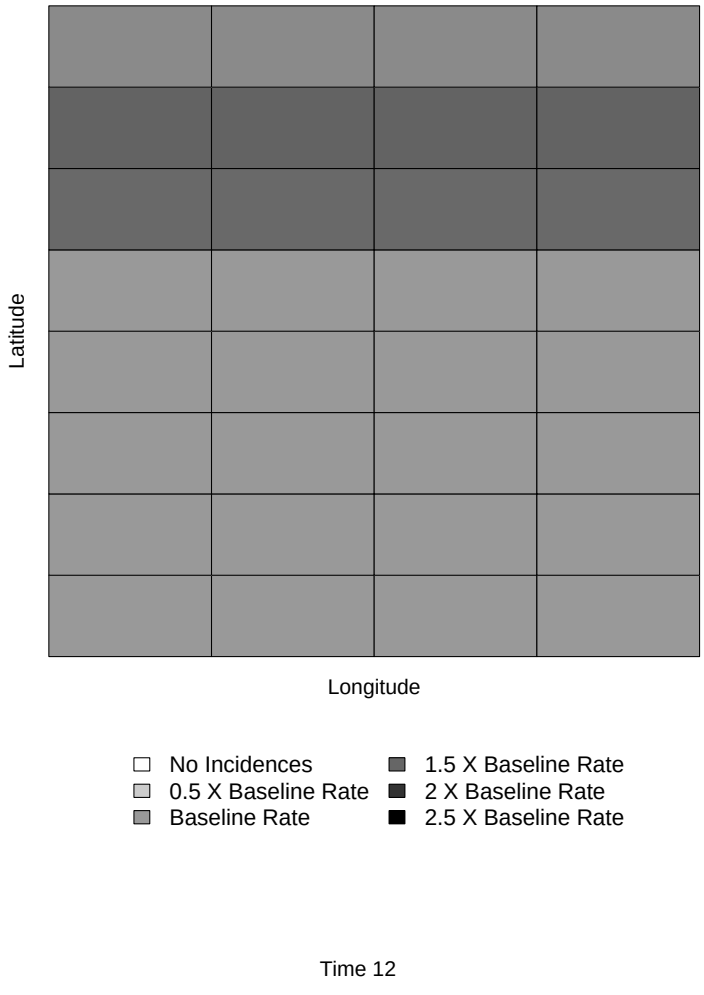


Figure 3.16: Ratios of estimated incidence rates to baseline for the first demonstration using the AIC reduced model

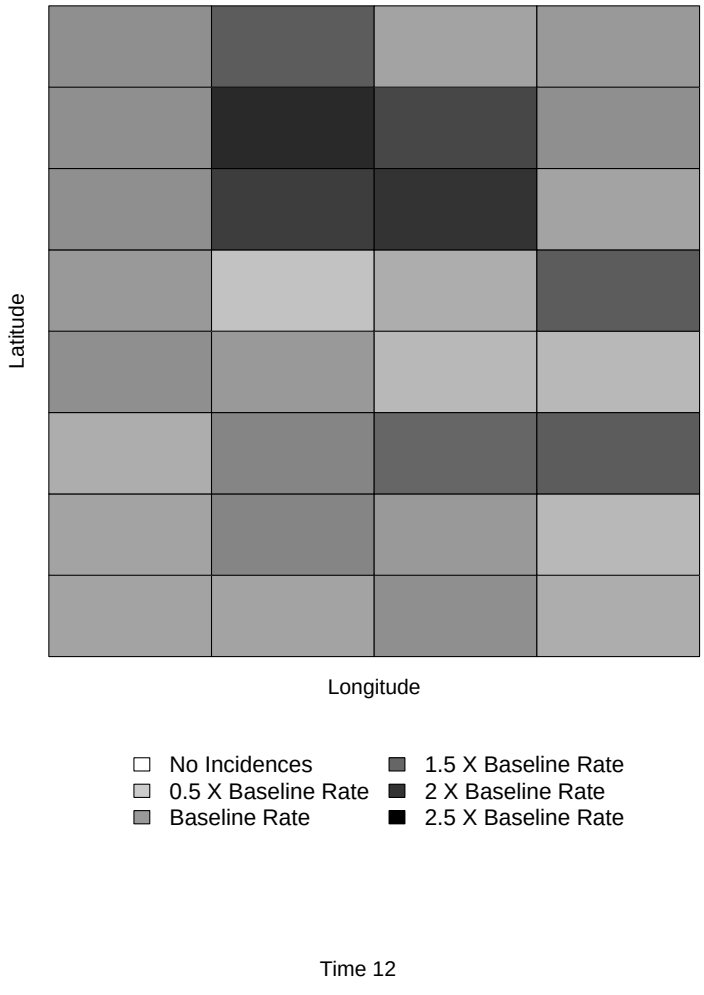


Figure 3.17: Ratios of estimated incidence rates to baseline for the first demonstration using the full model

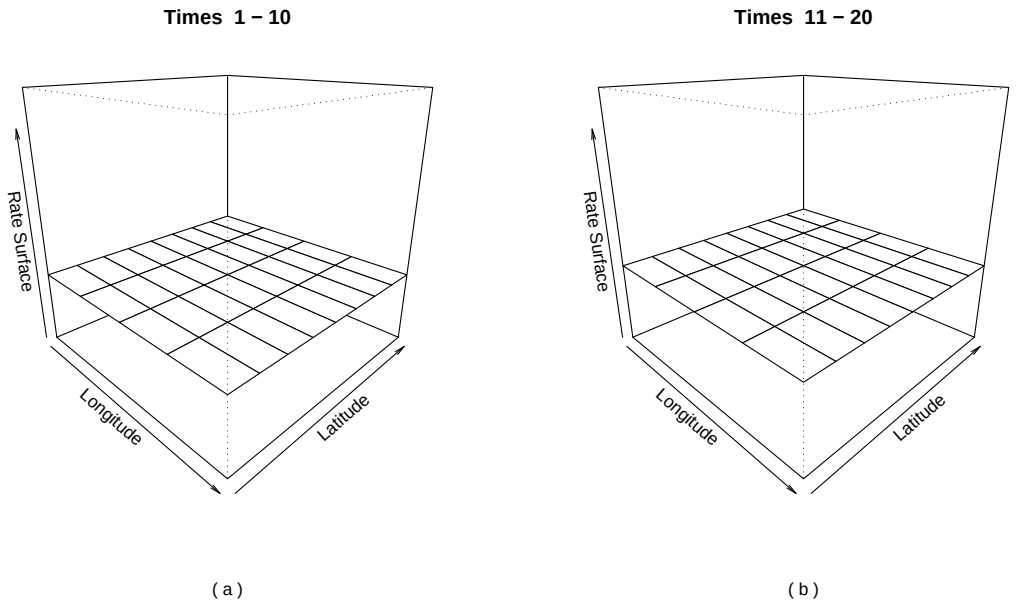


Figure 3.18: Incidence rate surfaces for the second demonstration: (a) Incidence rate surface for time intervals 1 to 10; (b) Incidence rate surface for time intervals 11 to 20

MEWMA, and Weighted χ^2 control charts using the Poisson regression model with the canonical link function all signaled based on the same simulated data. The Weighted χ^2 chart signaled at time interval 15, which was the same time interval where this chart signaled when using the identity link model. The chi-square and MEWMA control charts both signaled at time interval 17 when the canonical link model was used.

As in the first demonstration, the SWS, AIC, and full model diagnostics were used to estimate the changes in the incidence rate surface when the first signal occurred at time interval 15, based on the identity link model. Shaded maps of the estimated changes in the incidence rate surface from baseline over the wavelet domain are shown in Figures 3.22, 3.23, and 3.24, for the SWS, AIC, and full model diagnostics, respectively. In the scenario considered for this demonstration, the disease cluster covers the entire region and is estimated adequately by the intercept in the models. Therefore, there is not an over-smoothing issue with the SWS and AIC reduced models as there was in the previous demonstration. Using all three diagnostic plots, there is an indication that the disease cluster covers the entire region. The SWS diagnostic plot shows an estimated

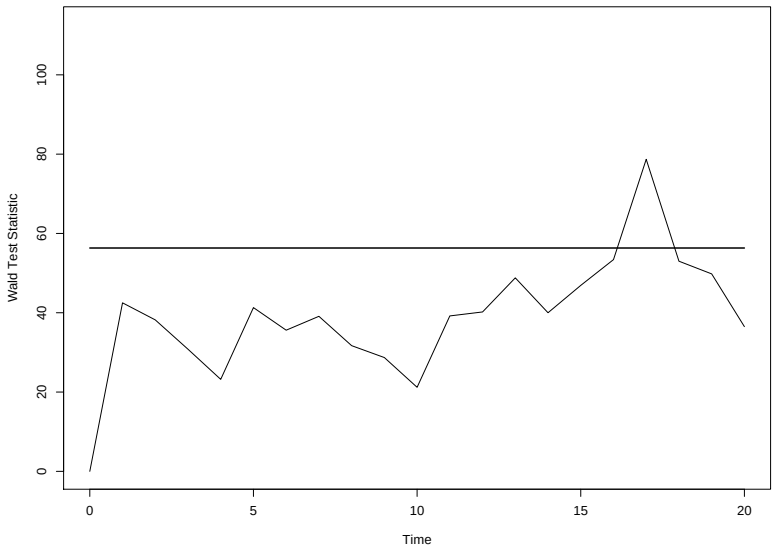


Figure 3.19: Chi-square control chart for the second demonstration

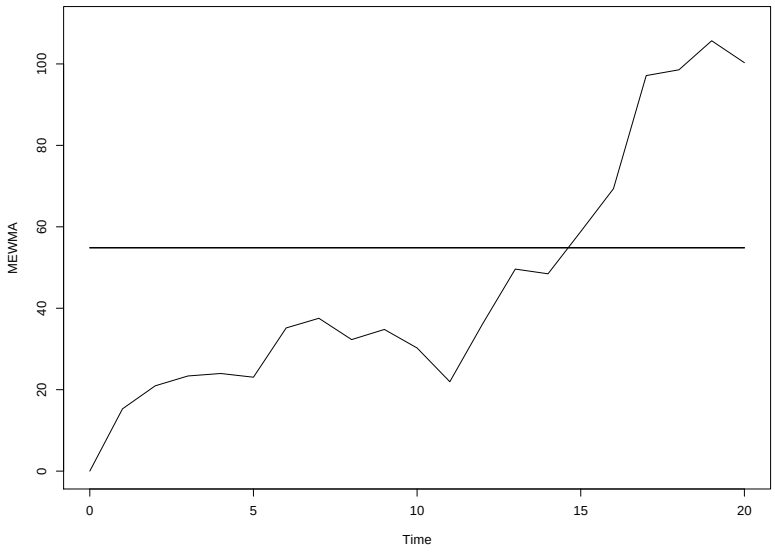


Figure 3.20: MEWMA control chart for the second demonstration

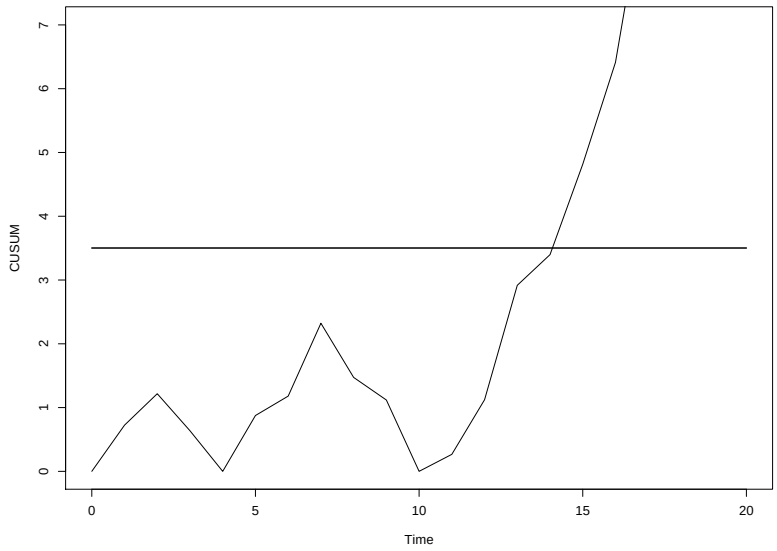


Figure 3.21: Weighted χ^2 control chart for the second demonstration

increase in the incidence rate from baseline in all 32 subregions. The AIC diagnostic plot shows increased incidence rate estimates in the majority of the subregions, but shows 8 subregions in the southern part of the region where the estimated incidence is slightly lower than baseline. The full model diagnostic plot also shows an estimated increase in the incidence rates from baseline for a majority of the subregions.

The third demonstration was performed to show that the control charts in Section 3.3.3 can be used to detect a cluster in one area of the region, when the mean incidence rate over the entire geographical region remains the same. In this case, the incidence rate increased in four of the 32 subregions, while the incidence rate simultaneously decreased in four other subregions. For time intervals 1 to 10, the incidence rate stayed at a baseline value of 10 per 100,000 residents in all of the subregions. For time intervals 11 to 20, the incidence rate increased to 12.739 per 100,000 residents in four adjacent subregions in the northern half of the geographical region and decreased to 7.261 per 100,000 residents in four adjacent subregions in the southern half, while the incidence rate in all other subregions remained at 10 per 100,000 residents. The incidence rate surfaces for the time intervals are shown in Figure 3.25. In this case, the

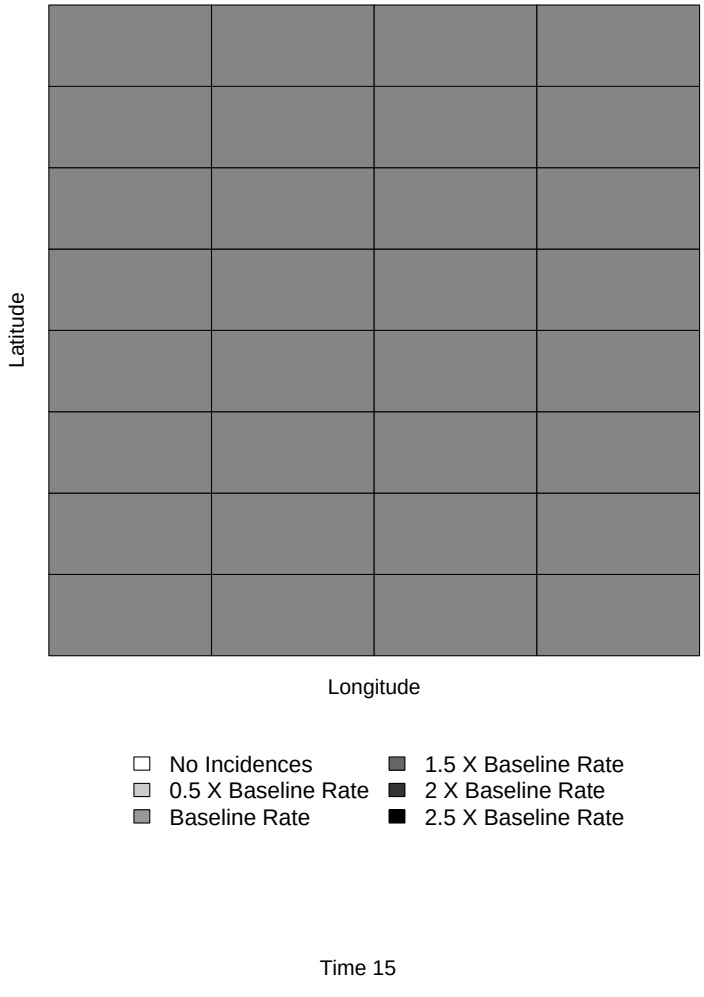


Figure 3.22: Ratios of estimated incidence rates to baseline for the second demonstration using the SWS reduced model

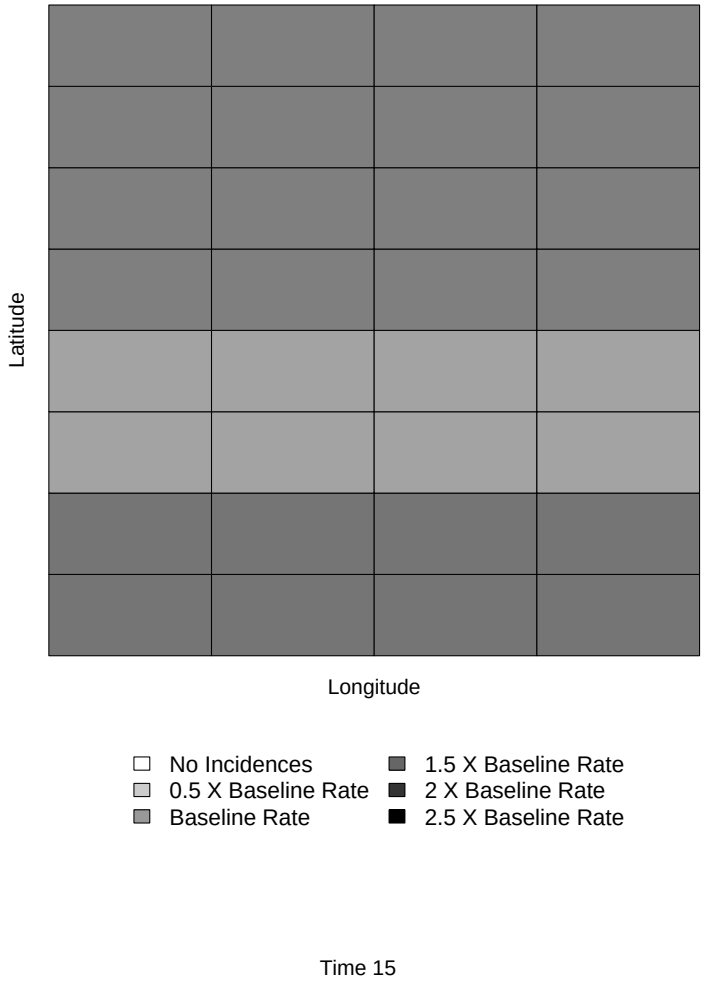


Figure 3.23: Ratios of estimated incidence rates to baseline for the second demonstration using the AIC reduced model

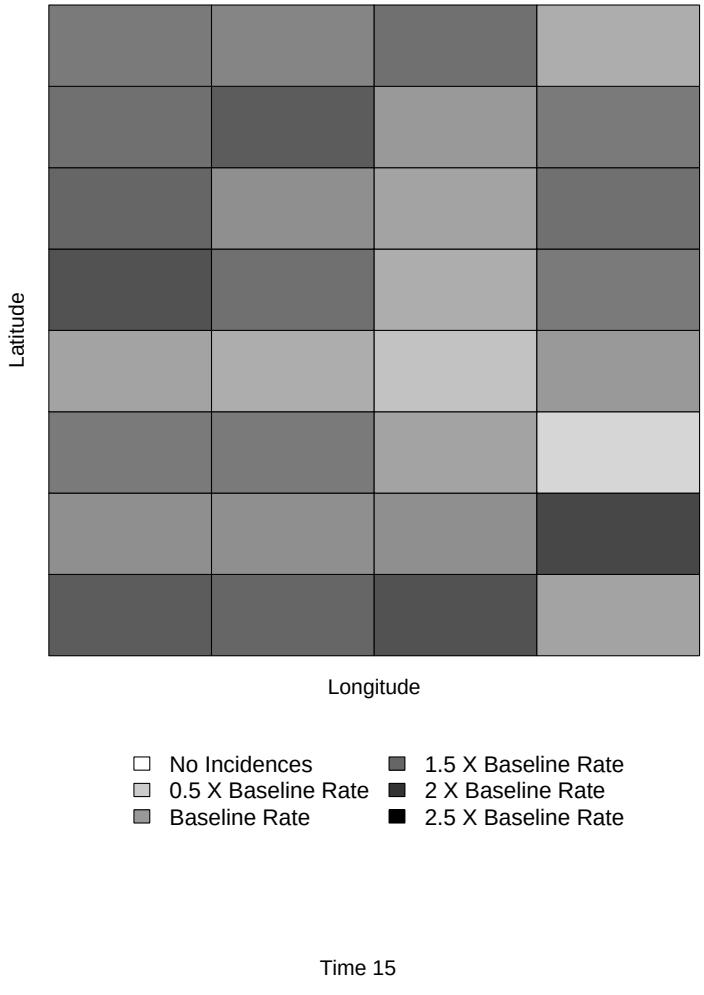


Figure 3.24: Ratios of estimated incidence rates to baseline for the second demonstration using the full model

change in the surface led to a shift in the coefficients that correspond to the incidence rates in the four northern subregions and four southern subregions, but there was no change in the mean incidence rate over all regions.

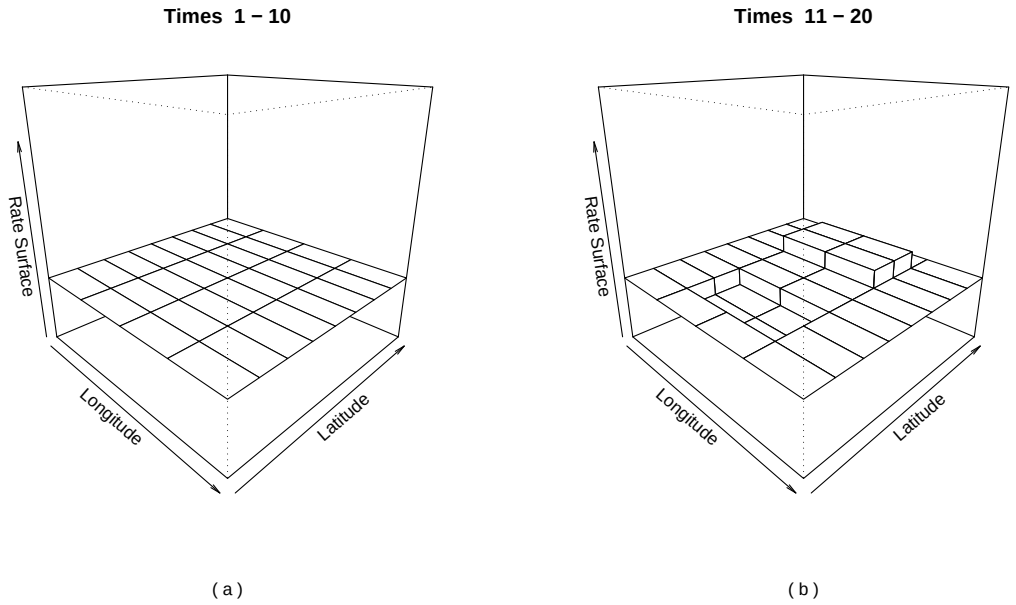


Figure 3.25: Incidence rate surfaces for the third demonstration: (a) Incidence rate surface for time intervals 1 to 10; (b) Incidence rate surface for time intervals 11 to 20

This third scenario shows the ability of the wavelet-based monitoring method to detect a cluster in one part of the region, while the incidence rate is decreasing in another part of the region and the overall incidence rate remains the same. The control charts used in this case were the same charts used for the first two demonstrations. Chi-square, MEWMA, and Weighted χ^2 control charts, based on randomly generated Poisson incidence counts and the use of the identity link model, are given in Figures 3.26, 3.27, and 3.28, respectively. The MEWMA and Weighted χ^2 control charts signaled for time intervals 15 and 16, respectively, indicating a change in the incidence rate surface. The chi-square control chart did not signal for any time interval. When using these control charts with the canonical link model on the same simulated data, the MEWMA and Weighted χ^2 control charts signaled for the same time intervals that they did when using the identity link model. The chi-square chart did not signal when

using the canonical link model. This is consistent with the out-of-control performance results in Section 4.2.2.1, which show that the chi-square control chart does not perform well in this type of scenario when using these two models.

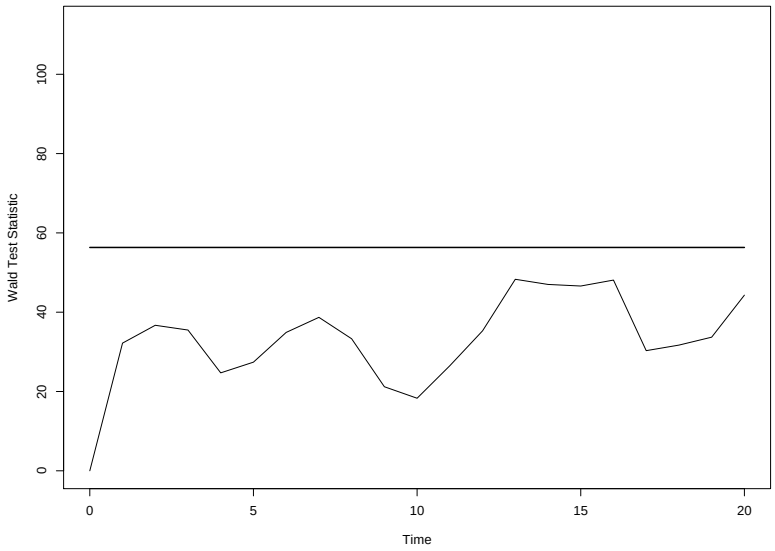


Figure 3.26: Chi-square control chart for the third demonstration

In this third demonstration, the SWS, AIC, and full model diagnostics based on the identity link model were again used to estimate the changes in the incidence rate surface when the first signal occurred for time interval 15. The shaded maps of the estimated changes in the incidence rates from baseline over the wavelet domain using the SWS, AIC, and full model diagnostics are shown in Figures 3.29,3.30, and 3.31, respectively. In this demonstration, the MEWMA control chart was the only chart that signaled at time interval 15, which was the first time interval a signal occurred. This indicates that the MEWMA control chart signaled based on observations obtained from previous time intervals as well as the current time interval. Therefore, the diagnostic plots are not optimal for identifying the areas with increased and decreased incidence rates in this demonstration as discussed in Section 3.3.4. Based on all three diagnosic plots, it is clear that there is an area with increased incidence rates in the northern part of the region, but the plots suggest that this cluster is only comprised of two

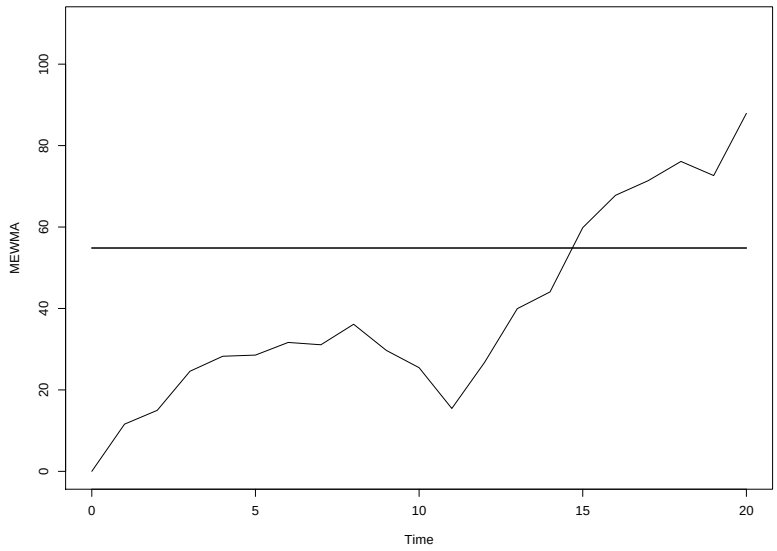


Figure 3.27: MEWMA control chart for the third demonstration

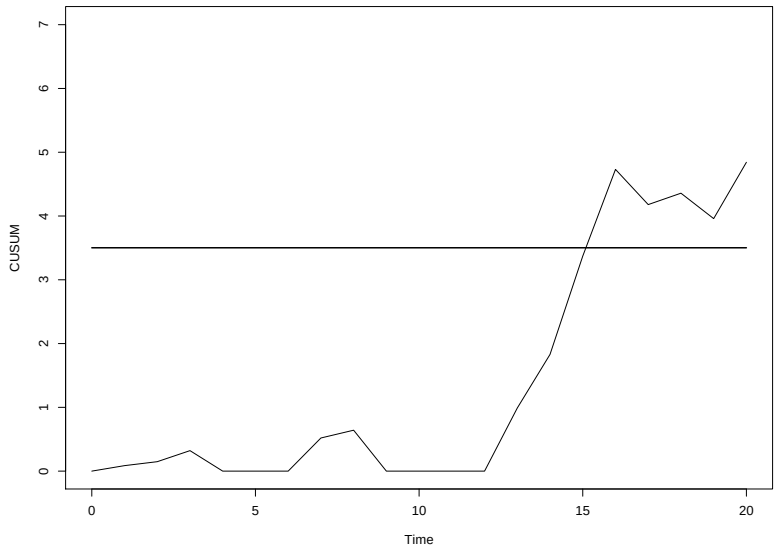


Figure 3.28: Weighted χ^2 control chart for the third demonstration

subregions instead of the four subregions known to be a part of this cluster. All three plots also show that the four center subregions in the southern half of the region have a decreased incidence rate, as they should, but these plots also indicate that several other subregions throughout the region have decreased incidence rates.

These three demonstrations show that the wavelet-based monitoring method is appropriately quantifying changes in the incidence rate surface and that the chi-square, MEWMA, and Weighted χ^2 control charts are effective for detecting different forms of clustering over a geographical region. It is important to note, however, that while the shifts in the mean incidence counts in these demonstrations are not unrealistic, they are somewhat large. One of the goals in developing this wavelet-based method is for the method to have the ability to detect small shifts in the incidence rate surface. The evaluation of this method, based on the in-control and out-of-control ARL performance presented in Section 4.2, shows that these control charts are also useful for detecting changes in incidence rate surfaces that are smaller in magnitude.

3.5 Summary and Discussion

The demonstrations of the wavelet-based monitoring method presented in Section 3.4 show that this method is useful for detecting clusters of disease prospectively when aggregated spatial and temporal incidence count data are available, and provided a reasonable mapping of the geographical region to the wavelet domain can be found. One aspect of the wavelet-based method that makes it useful is its flexibility. When implementing this method, one can choose between several different combinations of models, control charts, and diagnostic tools to suit the specific application. The performance of the control charts, which will be discussed in Section 4.2, varies based on the model selected and baseline incidence rate. Therefore, the practitioner can choose the model and control chart combination that has good properties for their specific application.

There are several other features of the wavelet-based surveillance method that also make it useful in disease monitoring applications. One of these features is the ability of the method to detect clusters of different size and shape within a geographical region, due to the multiresolution of wavelets. This was shown in the demonstrations presented

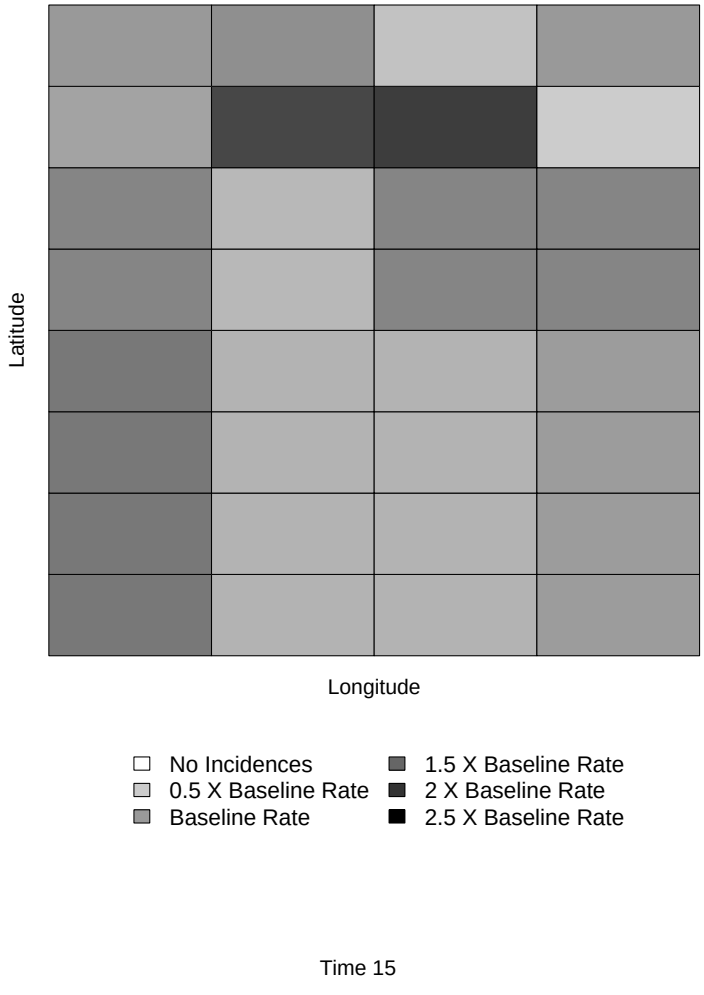


Figure 3.29: Ratios of estimated incidence rates to baseline for the third demonstration using the SWS reduced model

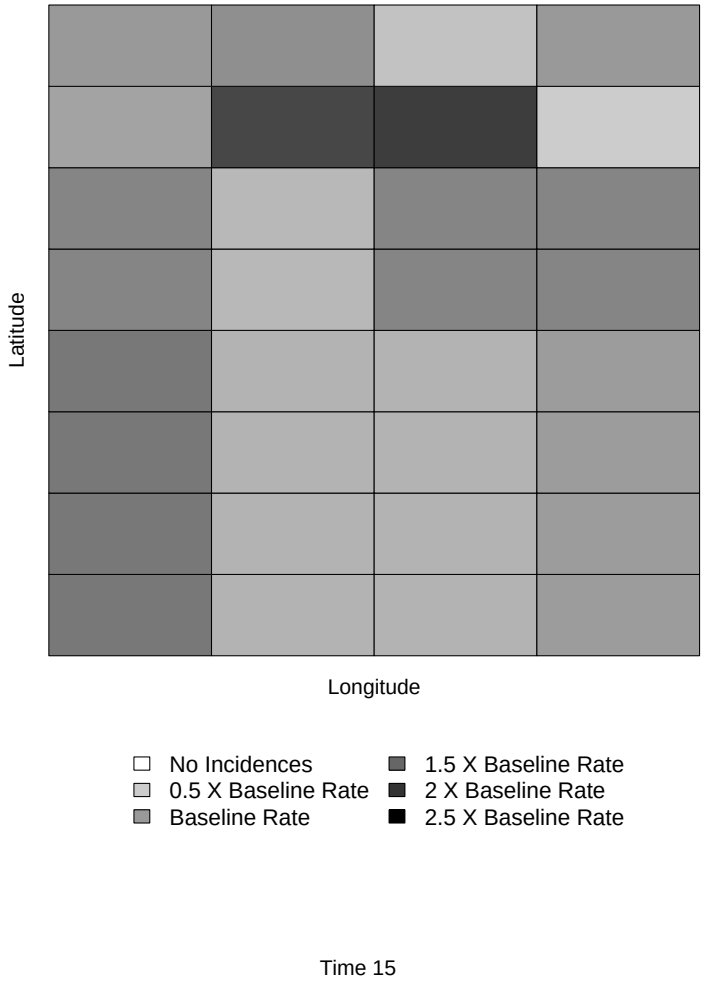


Figure 3.30: Ratios of estimated incidence rates to baseline for the third demonstration using the AIC reduced model

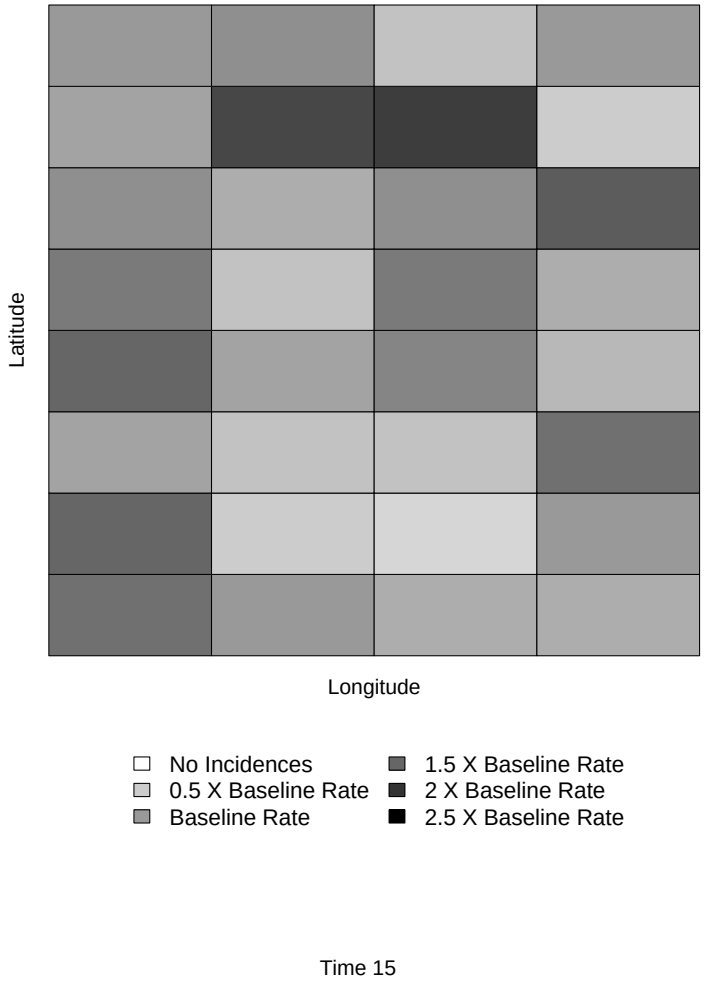


Figure 3.31: Ratios of estimated incidence rates to baseline for the third demonstration using the full model

in Section 3.4 through use of the control charts and diagnostics. The multiresolution was also important in the development of the weighted χ^2 statistic, which can be used in the Weighted χ^2 control chart to detect disease clusters of a specific size. Another feature of this method is that it does not require multiple testing, like the methods of Raubertas (1989) and Kleinman (2005). In many methods, a test is done in each subregion at each time period and these tests can accumulate quickly, especially if there are a large number of subregions. In the wavelet-based method, there is only one test done for cluster detection at each time period, which is the value plotted on the control chart. An analysis of the entire region is only done, through the use of diagnostics, when the control chart signals and it is likely that a true cluster is present.

The wavelet-based surveillance method is very similar to the methods of Rogerson and Yamada (2004) and Kleinman (2005). The wavelet-based method has an advantage over both of these methods based on how it adjusts for population. This method also has an advantage over the method of Rogerson and Yamada (2004) based on how it incorporates covariate information. Both the wavelet-based method and the method of Kleinman (2005) use a Poisson regression model to estimate incidence counts. In the wavelet-based method an offset is used to adjust for changes in the population and covariates over time in cases when the canonical link model is used, and the baseline incidence counts are adjusted at each time point to adjust for changes in the population and covariates when the identity link model is used. Kleinman (2005), however, makes no adjustment for population, which means a cluster may be indicated only due to an increase or shift in the population. The method of Rogerson and Yamada (2004) has a similar problem because it accounts for differing population sizes and covariates within each subregion by adjusting the baseline incidence counts only once before monitoring begins. Since these baseline counts are compared to the observed counts for each time interval, when the population or covariates change within subregions over time, an increase in the counts may be detected and interpreted as a disease cluster when the counts only increased due to a change in the population or a change in the covariates.

The wavelet-based method is also very similar to the method of Kulldorff (2001), but the wavelet-based method has some definite advantages. In the scan method of Kulldorff (2001), a prospective space-time scan statistic is calculated each time new data are available, based on multiple space-time scanning windows. This statistic indi-

cates only the most likely cluster, which makes it impossible for the method to detect multiple clusters. Kulldorff (2001) suggests looking at the subregions associated with several of the largest likelihood ratio statistics calculated for each scanning window, instead of only the subregions associated with the maximum of these likelihood ratios. This can allow for the detection of multiple clusters, but in most cases it will only identify additional subregions close to the most likely cluster. In contrast, the wavelet-based method can detect multiple clusters easily. The range of cluster sizes that can be detected using the method of Kulldorff (2001) is also smaller than that of the wavelet-based method. The spatial part of the scanning window, used to calculate the prospective space-time scan statistic, only increases to a size that covers 50% of the population in the geographical region. This does not allow for the detection of very large clusters or an increase in the mean incidence rate over the entire region of interest. Also, since the spatial part of the scanning window is circular, it is limited in the shape of clusters that it can detect, which is not an issue with this wavelet-based method. The method of Kuldorff (2005) could potential be extended, however, to detect non-circular clusters based on the work of Tango and Takahashi (2005). Tango and Takahashi (2005) extended the spatial scan statistic by searching over irregularly-shaped windows, not just circular windows. Another drawback of the method of Kulldorff (2001) is that it assumes that all subregions have the same baseline incidence rate per person. The wavelet-based method does not require such a strict assumption and allows for different baseline incidence rates within each subregion.

Chapter 4

Evaluation of the Wavelet-Based Method for the Prospective Monitoring of Disease Occurrences

4.1 Introduction

Before implementing a disease surveillance method, it is important to evaluate its performance in different scenarios to understand how the method is expected to perform in the planned application. It is also important to compare the performance of surveillance methods to other related methods to determine if another method may be more suitable for a specific application. When considering spatio-temporal surveillance methods, computing the average run length (ARL) performance of the surveillance methods under different monitoring scenarios is an effective way to compare methods overall, and to compare different parameter settings for a specific method. Based on these types of performance comparisons, one can select a surveillance method, along with the parameters of the method, that should be effective for detecting clusters of disease when they are truly present with few false signals. These types of comparisons are seldom done when new spatio-temporal disease surveillance methods are developed. Rogerson and Yamada (2004), Sonesson (2007), Joner *et al.* (2008), Zhou and Lawson (2008), Neill (2009), and Neill and Cooper (2009) are among those who have examined the ARL characteristics of their spatio-temporal surveillance methods.

To adequately evaluate the wavelet-based disease surveillance method, the ARL performance will be determined for the control charts that have been developed for

this method. Additional control charts that are representative of other spatio-temporal surveillance methods, but applied within the wavelet-based method, will also be evaluated based on their ARL performance. These additional control charts are discussed in Section 4.2. Both the in-control and out-of-control ARL performance of the control charts will be determined under different scenarios. The in-control ARL performance of the control charts is discussed in Section 4.2.1. In-control ARLs can be used to evaluate the false alarm behavior of the control charts compared to the false alarm behavior expected when designing the charts. The in-control ARL performance also indicates how well the distribution of the statistics used for monitoring are approximated. The out-of-control ARL performance of the control charts is discussed in Section 4.2.2. Out-of-control ARLs can be used to evaluate the control charts in terms of their ability to detect disease clusters when these clusters are truly present.

In addition to evaluating the ARL performance of a disease surveillance method, it is important to demonstrate and evaluate how it performs in a real surveillance situation. This can be accomplished through a case study by implementing the surveillance method using real-world data and evaluating its performance based on the expected outcome. The data available from the SEER program on female respiratory lung cancer incidences in New Mexico will be used for this purpose. This will provide a realistic example of how the method should be implemented because there are 19 years of monthly incidence count data available, and the yearly population count within each subregion is also included. This particular data set was chosen because it provides enough data to initially determine a baseline rate in each county and the remainder of the data can then be used to demonstrate the wavelet-based disease surveillance method. The case study examining the development of clusters of respiratory lung cancer in New Mexico is discussed in Section 4.3.

4.2 Control Chart Comparison based on ARL Performance

A comparison of ARLs for the control charts used in the wavelet-based surveillance method was performed to determine the false alarms rates of the method when no

disease clusters are present and how well the method detects clusters when they are truly present. The wavelet-based charts included in this comparison are the chi-square, multivariate exponentially weighted moving average (MEWMA), and Weighted χ^2 control charts described in Section 3.3.3. Two other control charting methods were also included in this comparison to show how this method performs when an alternative approach to monitoring is used. The alternative control charts used were one-sided and two-sided univariate cumulative sum (CUSUM) charts that were used to monitor each subregion individually. The one-sided charts were designed to detect only an increase in the disease incidence rate within the subregions, which would indicate the formation of a disease cluster. The two-sided charts were designed to detect either an increase or decrease in the incidence rate within each subregion. This two-sided approach corresponds more closely to the use of the chi-square, MEWMA, and Weighted χ^2 control charts since they can detect both increases and decreases in the incidence rates. These two additional methods were chosen for comparison because similar methods have been considered for spatio-temporal disease surveillance by other researchers, such as Raubertas (1989) and Rogerson and Yamada (2004).

The CUSUM values plotted on the one-sided and two-sided subregion control charts are based on statistics calculated from the model used in the wavelet-based surveillance method. When the Poisson regression model using the canonical link function is used, the statistic calculated for each time interval i and subregion s is

$$R_s(i) = \left(\boldsymbol{a}_s^T\hat{\boldsymbol{\beta}}(i) - ln\left(\lambda_{0_s}\right)\right)\left[\boldsymbol{a}_s^T\left(\boldsymbol{X}^T\hat{\boldsymbol{F}}(i)\boldsymbol{X}\right)^{-1}\boldsymbol{a}_s\right]^{-1}, \tag{4.1}$$

where $\boldsymbol{a}_s$ is a vector representing row s of the X matrix, which corresponds to wavelet coefficients for subregion s, $\hat{\boldsymbol{\beta}}(i)$ is the estimated wavelet coefficient vector for observation i, λ_{0_s} is the baseline incidence rate for subregion s, and $\hat{\boldsymbol{F}}(i)$ is a $N_{wd} \times N_{wd}$ diagonal matrix with diagonal elements equal to $e^{ln[\boldsymbol{N}(i)]+ln[\boldsymbol{C}(i)]+\boldsymbol{X}\hat{\boldsymbol{\beta}}(i)}$. When the Poisson regression model using the identity link function is used, the statistic calculated for each time interval i and subregion s is

$$R_s(i) \;=\; \left(\boldsymbol{a}_s^T\hat{\boldsymbol{\beta}}(i) - p_s(i)c_s(i)\lambda_{0_s}\right)\left[\boldsymbol{a}_s^T\left(\boldsymbol{X}^T\boldsymbol{\Sigma}_0(i)^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{a}_s\right]^{-1}, \tag{4.2}$$

where $p_s(i)$ is the population in subregion s for time interval i, $c_s(i)$ is the covariate adjustment in subregion s for time interval i, $\boldsymbol{\Sigma}_0(i)$ is the $N_{wd} \times N_{wd}$ covariance matrix of the baseline incidence counts for observation i, and all other variables have the same definitions given for the canonical link model. The statistic $R_s(i)$ is assumed to have an approximate standard normal distribution. The validity of this assumption will be discussed in Section 4.2.1.

Based on the assumption of normality, the values plotted on the one-sided CUSUM control chart for each subregion s are

$$C_s(j)_{upper} = max\left(0, R_s(j) - \frac{1}{2} + C_s(j-1)_{upper}\right), \tag{4.3}$$

for $j = 1, 2, \ldots, i$, where i is the current observation and $C_s(0)_{upper} = 0$ for all subregions. The one-sided chart for subregion s signals when $C_s(j)_{upper} \geq h$. The values in equation (4.3) are also plotted on the two-sided CUSUM control chart for each subregion s along with the values

$$C_s(j)_{lower} = min\left(0, R_s(j) + \frac{1}{2} + C_s(j-1)_{lower}\right), \tag{4.4}$$

for $j = 1, 2, \ldots, i$, where i is the current observation and $C_s(0)_{lower} = 0$ for all subregions. The two-sided chart for subregion s signals when $C_s(j)_{upper} \geq h$ or when $C_s(j)_{lower} \leq -h$. When implementing the one-sided and two-sided subregion CUSUM methods, the value of h is chosen to achieve an approximate specified in-control ARL for all control charts combined by assuming that the values of $R_s(j)$ follow a standard

normal distribution or to achieve an exact specified in-control ARL for all control charts combined by simulation. When using the one-sided or two-sided CUSUM approach, if at least one of the CUSUM charts for an individual region signals, it is considered a signal that there has been a change in the incidence rate surface and that a disease cluster has formed. Therefore, the univariate CUSUM control charting methods can be thought of as a system of control charts working together. The one-sided and two-sided CUSUM methods will be referred to as the one-sided and two-sided subregion control charts, respectively.

4.2.1 In-Control ARL Performance

Understanding the in-control ARL performance of control charts used in monitoring systems is important for two reasons. The in-control ARL performance provides the expected false alarm rate of the chart and indicates how often the chart will signal when there are no changes in the parameters being monitored. When implementing control charts it is important for the in-control ARL to be large so that false signals occur infrequently; however, this value should not be so large that it prohibits the chart from detecting true changes in the parameter values when they occur. The in-control ARL of a control chart is also important for assessing the validity of the assumptions made when implementing a chart. Specifically, this value can be used to determine how well the distribution assumed for the monitored parameters approximates the true distribution of the parameters. If the true in-control ARL is close to the nominal in-control ARL chosen when determining the control limit for the chart, then the approximate distribution estimates the true distribution of the monitored parameters well. For these reasons, the in-control ARL performance of the control charts within the wavelet-based disease surveillance method was investigated.

In the wavelet-based disease surveillance method, the in-control ARL performance for each control chart is influenced by several factors. One factor is the number of subregions in the wavelet domain, N_{wd}. Changes in this value change the number of wavelet coefficients being monitored. The in-control ARL performance will also be affected by the baseline incidence counts, since the values of the counts impact the approximate distribution of the monitored coefficients. In addition, the in-control

ARL performance will change depending on the Poisson regression model used when the monitoring method is implemented. To assess the impact of these factors on in-control ARL performance of the wavelet-based method, simulations were performed for different combinations of these factors. For each control chart, the in-control ARL was estimated for regions with 16 and 32 subregions, baseline values of 1, 5, 10, 20, 50, and 100 per 100,000 residents, and for both the canonical link and identity link models. The control limits for each chart were determined so that each chart had a nominal or expected in-control ARL of 200, based on the approximate distribution of the monitored coefficients. The control limits used for this evaluation are shown in Table 4.1. These control limits were computed using the anygeth.exe program available from Hawkins and Olwell (1998).

Table 4.1: Control limits for a nominal in-control ARL of 200 for each control chart used to evaluate the in-control ARL performance of the wavelet-based disease surveillance method

	Control Chart				
Subregions	Chi-Square	MEWMA	Weighted χ^2	One-Sided Subregion	Two-Sided Subregion
16	34.37	33.03	3.502	6.222	6.914
32	56.33	54.85	3.502	6.914	7.605

To simulate the in-control state for each combination of number of subregions and baseline rate over time, the population in each subregion was fixed at 100,000 for each time interval and no covariate adjustments were used. Then, for each time interval, a random independent observation consisting of counts for each subregion was generated from the Poisson distribution corresponding to the specified baseline rate. After each observation was obtained, values to be plotted on each control chart were calculated for both models. These values were then compared to the control limits for each corresponding control chart. This process was repeated for each control chart until the chart signaled and one run length was produced. To estimate the in-control ARL, 20,000 run lengths were obtained under these same conditions for each control chart and then averaged.

Tables 4.2 and 4.3 contain the in-control ARL simulation results for the canonical and identity link models, respectively. The standard errors, SE, are also provided

to show the precision of the ARL estimates. Figures 4.1 and 4.2 show how the in-control ARLs compare to the nominal ARL value of 200 for all control charts for the canonical and identity link models, respectively. Generally, these results show that for both models and both numbers of subregions, as the incidence count increases the true in-control ARL gets closer to the nominal ARL of 200 for each control chart. This indicates that as the baseline count increases, the approximate distribution assumed for the wavelet coefficients monitored in each control chart more closely represent the true distributions. The in-control ARLs for the two-sided subregion control chart, when either model is used, begin with values below the nominal ARL when the baseline count is 1, but increase to values above 200 as the baseline count increases. This phenomenon also occurs for the Weighted χ^2 chart when the identity link model is used and the number of subregions is 16. Further in-control ARL simulations would need to be performed to determine if these ARLs will remain above or converge to 200 as the baseline count increases beyond 100, and to determine if this phenomenon occurs with other control charts at higher baseline counts. In most cases the incidence rates monitored with this method would be low, however, so the in-control ARL performance for a large number of baseline counts is not as important as the performance at lower counts.

The in-control ARL performance for the control charts when there are 16 versus 32 subregions differs depending on the Poisson regression model used. When the canonical link model is used, the chi-square and MEWMA control charts have better performance, meaning they have in-control ARLs that approach 200 more quickly as the baseline count increases, when there are 16 subregions as opposed to 32 subregions. The opposite is true for the Weighted χ^2 control chart. Both the one-sided and two-sided subregion control charts have comparable performance for 16 and 32 subregions. When the identity link model is used, in most cases, the in-control ARL performance for each control chart is comparable when there are 16 versus 32 subregions. The Weighted χ^2 and one-sided subregion charts are the exceptions. The Weighted χ^2 control chart has better performance for baseline counts up to 20 when there are 16 subregions and better performance for baseline counts 50 and 100 when there are 32 subregions. The one-sided region control chart performs better for baseline counts up

Table 4.2: In-control ARL estimates using the Poisson regression canonical link model for two different numbers of subregions and for six baseline rates per 100,000 residents

		Control Chart				
Subregions	Baseline Rate	Chi-Square *ARL* (*SE*)	MEWMA *ARL* (*SE*)	Weighted χ^2 *ARL* (*SE*)	One-Sided Subregion *ARL* (*SE*)	Two-Sided Subregion *ARL* (*SE*)
16	1	339.3 (2.42)	2.0 (0.00)	>10,000 (N/A)	39.0 (0.21)	55.9 (0.31)
	5	235.8 (1.67)	9.5 (0.05)	2676.5 (18.80)	121.9 (0.81)	217.4 (1.47)
	10	211.1 (1.49)	50.3 (0.31)	561.6 (3.91)	140.6 (0.93)	223.0 (1.50)
	20	204.9 (1.43)	109.9 (0.74)	333.9 (2.35)	151.1 (1.02)	217.7 (1.47)
	50	204.0 (1.44)	163.3 (1.11)	253.4 (1.76)	166.7 (1.12)	210.8 (1.41)
	100	200.7 (1.43)	181.8 (1.25)	231.4 (1.60)	179.4 (1.22)	215.1 (1.43)
32	1	1047.2 (7.37)	2.0 (0.00)	>10,000 (N/A)	34.2 (0.16)	46.7 (0.23)
	5	330.9 (2.34)	5.8 (0.03)	2548.9 (18.04)	116.1 (0.77)	213.2 (1.44)
	10	244.1 (1.72)	36.0 (0.21)	410.0 (2.88)	135.9 (0.89)	219.1 (1.46)
	20	220.0 (1.54)	91.4 (0.59)	283.5 (1.96)	148.9 (0.98)	214.8 (1.43)
	50	204.2 (1.44)	149.5 (1.01)	230.8 (1.62)	164.0 (1.09)	211.9 (1.42)
	100	204.7 (1.45)	174.6 (1.20)	217.6 (1.51)	179.2 (1.20)	218.0 (1.45)

Table 4.3: In-control ARL estimates using the Poisson regression identity link model for two different numbers of subregions and for six baseline rates per 100,000 residents

		Control Chart				
Subregions	Baseline Rate	Chi-Square *ARL* (*SE*)	MEWMA *ARL* (*SE*)	Weighted χ^2 *ARL* (*SE*)	One-Sided Subregion *ARL* (*SE*)	Two-Sided Subregion *ARL* (*SE*)
16	1	46.5 (0.33)	126.3 (0.86)	105.0 (0.73)	60.1 (0.39)	87.3 (0.58)
	5	111.4 (0.79)	175.7 (1.18)	177.5 (1.22)	104.3 (0.69)	169.0 (1.13)
	10	139.5 (0.99)	187.5 (1.28)	194.1 (1.34)	128.2 (0.84)	185.2 (1.23)
	20	164.4 (1.15)	191.2 (1.32)	202.2 (1.41)	142.9 (0.97)	197.5 (1.33)
	50	184.8 (1.30)	196.7 (1.34)	208.7 (1.44)	162.4 (1.09)	203.2 (1.36)
	100	190.4 (1.35)	200.2 (1.38)	210.4 (1.45)	179.4 (1.22)	215.1 (1.43)
32	1	45.8 (0.32)	128.4 (0.87)	84.0 (0.58)	47.3 (0.29)	96.6 (0.63)
	5	112.3 (0.79)	177.5 (1.21)	159.0 (1.11)	100.4 (0.65)	158.6 (1.05)
	10	147.1 (1.05)	189.3 (1.29)	177.9 (1.22)	120.3 (0.78)	182.8 (1.21)
	20	166.5 (1.17)	195.2 (1.32)	190.5 (1.32)	139.6 (0.92)	194.6 (1.29)
	50	182.3 (1.29)	197.7 (1.34)	198.1 (1.38)	160.1 (1.06)	203.9 (1.37)
	100	192.6 (1.35)	201.0 (1.39)	201.4 (1.39)	179.2 (1.20)	218.0 (1.45)

to 10 when there are 16 subregions and has comparable performance for 16 and 32 subregions for baseline counts 20 through 100.

For the majority of the control charts, the in-control ARLs are well below the nominal value of 200 when the baseline counts are low and either the canonical or identity link models are used; however, the chi-square and Weighted χ^2 control charts have in-control ARLs much higher than 200 when the baseline counts are low and the canonical link model is used. In fact, when the baseline count is 1, the Weighted χ^2 control chart using the canonical link model has an in-control ARL so large that it could not be estimated exactly with the computing resources available. In cases where the true in-control ARL is below 200, the false alarm rate can be higher than what may be considered acceptable. When the true in-control ARL is well above 200, the false alarm rate will be low, but the charts may not have the ability to detect clusters of disease as they form. At low values of the baseline count, the control charts using the identity link model generally have in-control ARLs that are closer to the nominal in-control ARL of 200 when compared to the control charts using the canonical link model. For the identity link model, the MEWMA and Weighted χ^2 control charts exhibit the best performance when the baseline incidence counts are low. Therefore, if one wants to achieve good in-control ARL performance at low counts using the approximate distribution of the monitored wavelet coefficients, one of these control charts is recommended. If another control chart or model is preferred, or one wants a chart with an exact specified in-control ARL when the baseline rate is low, another option is to determine the control limit that will give an in-control ARL at the nominal value desired through the use of simulation.

4.2.2 Out-of-Control ARL Performance

The out-of-control ARL performance of control charts in any monitoring method should be considered when choosing or designing a control chart for a specific application, because the out-of-control ARL performance indicates how well the method will detect different shifts in the parameters being monitored. When selecting a control chart for a specific application, it is important that the out-of-control ARL is low, which indicates that the control chart will detect changes in the parameters being monitored

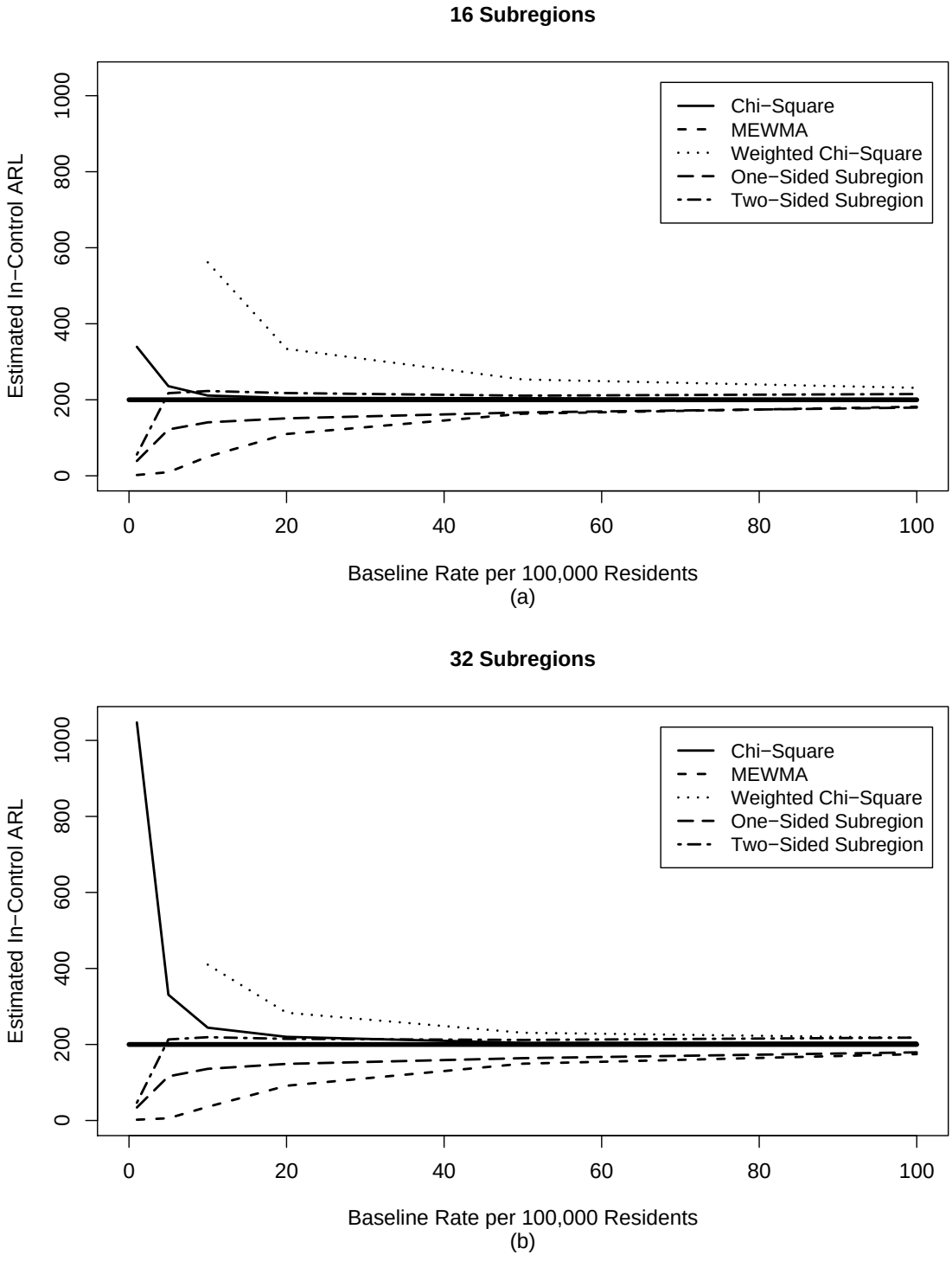
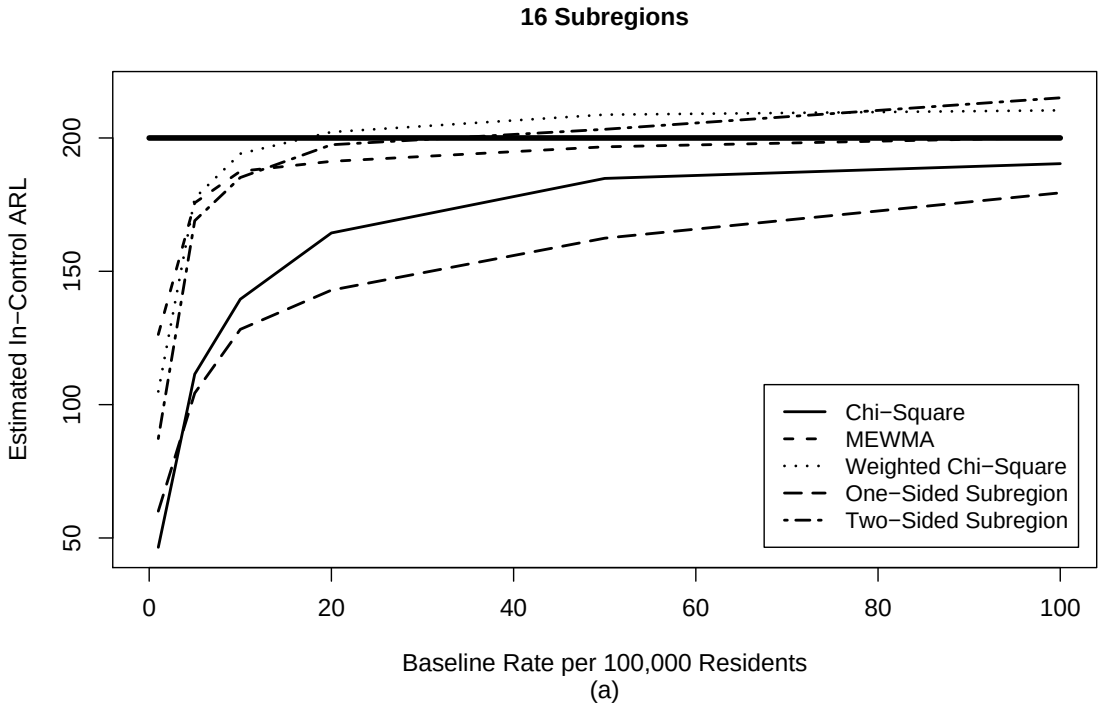


Figure 4.1: Estimated in-control ARL performance using the Poisson regression canonical link model for (a) 16 subregions; (b) 32 subregions



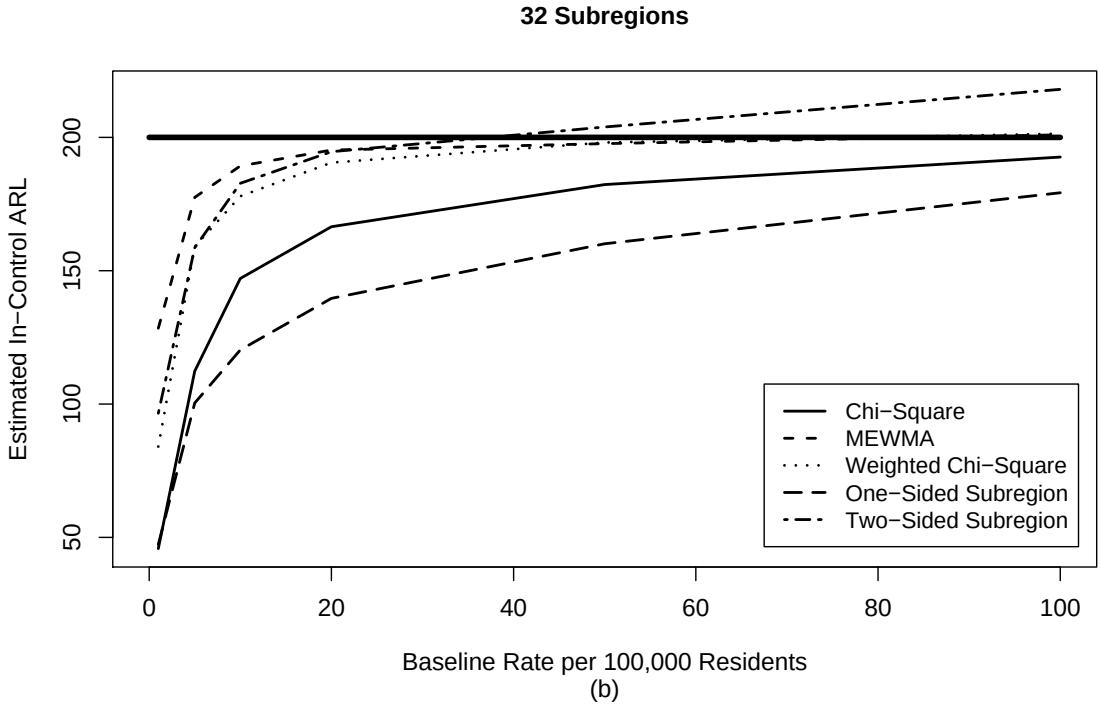


Figure 4.2: Estimated in-control ARL performance using the Poisson regression identity link model for (a) 16 subregions; (b) 32 subregions

quickly. In the wavelet-based disease surveillance method, a control chart that has a lower out-of-control ARL, when compared to other control charts for the same out-of-control scenario, will be able to detect clusters of disease more quickly on average. To compare the control charts that can be implemented using the wavelet-based method, the out-of-control ARL performance was evaluated for these charts by simulation.

In the wavelet-based method, the out-of-control ARL performance of each control chart is influenced by the same factors that have an impact on the in-control ARL performance. These factors include, the number of subregions, the baseline incidence counts, and the model used. In addition, the out-of-control ARL performance is affected by the magnitude and direction of the shift in the wavelet coefficient vector being monitored. In the wavelet-based method, differences in the direction of coefficient shifts can represent many different clustering scenarios. Specifically, these different coefficient shifts represent clusters of differing size, shape, and location. In the out-of-control ARL evaluation of the wavelet based method, the focus was on determining how the direction of the shift in coefficients influences the ARL performance. Changes in the magnitude of these shifts were not considered since one expects the pattern of control chart performance when comparing shifts in different directions to be similar at different magnitudes. At different magnitudes, the out-of-control ARLs are merely expected to increase or decrease as a whole.

To compare the out-of-control ARL performance for coefficient shifts in multiple directions, the magnitude of the coefficient shifts must remain constant. When evaluating out-of-control ARL performance of the wavelet-based method and the canonical link Poisson regression model was used, the magnitude of the change in the coefficients was made approximately the same in each case through the use of a noncentrality parameter. The noncentrality parameter was calculated using the baseline and out-of-control coefficient vectors. The form of the noncentrality parameter in this case was

$$\delta = \frac{1}{2}\left[\boldsymbol{\beta}_1 - \boldsymbol{\beta}(B)\right]^T \boldsymbol{\Sigma}\left(\mathbf{pc}\boldsymbol{\lambda}_1\right)^{-1}\left[\boldsymbol{\beta}_1 - \boldsymbol{\beta}(B)\right], \tag{4.5}$$

where

$$\boldsymbol{\Sigma}(\mathbf{pc}\boldsymbol{\lambda}_1)^{-1} = \boldsymbol{X}^T\boldsymbol{F}(\mathbf{pc}\boldsymbol{\lambda}_1)\boldsymbol{X}, \tag{4.6}$$

$\boldsymbol{\beta}_1$ is the out-of-control coefficient vector, $\boldsymbol{\beta}(B)$ is the baseline coefficient vector, $\mathbf{pc}\boldsymbol{\lambda}_1 = e^{ln(\mathbf{p})+ln(\mathbf{c})+\boldsymbol{X}\boldsymbol{\beta}_1}$, $\mathbf{p}$ is the vector of assumed population sizes for each subregion, $\mathbf{c}$ is the vector of assumed covariate adjustments for each subregion, and $\boldsymbol{F}$ is a $N_{wd} \times N_{wd}$ diagonal matrix with diagonal elements equal to the values in parentheses. This noncentrality parameter is comparable to the noncentrality parameter used to quantify shifts in a mean vector of exactly multivariate normal random variables, but in this case the covariance matrix is dependent on the coefficient vector, $\boldsymbol{\beta}_1$, because the incidence counts are assumed to follow independent Poisson processes. This noncentrality parameter corresponds to the approximate distribution of the shifted coefficient estimates and, as a result, only indicates the approximate magnitude of the shift in the coefficients. When the identity link Poisson regression model was used as part of the wavelet-based method, the noncentrality parameter in equation (3.42) was used to ensure the magnitude remained constant for different shifts in direction of the wavelet coefficient vector. In all simulations performed to evaluate out-of-control ARL performance of the wavelet-based method, the value of the noncentrality parameter, δ, was equal to 1.

In addition to making the magnitude consistent for each shift of the coefficient vector, the in-control ARL performance of each control chart in the wavelet-based method must be equivalent in order for their out-of-control ARL performance to be compared. This is achieved by finding the control limit for each control chart that produces the same value for the in-control ARL. Before simulations were performed to evaluate the out-of-control performance of the control charts in the wavelet-based method, a search was done to determine the control limits that produced an in-control ARL of approximately 200 for each chart based on the model used, the number of subregions, and the baseline incidence rate per 100,000 residents. The control limits, CL, determined by the search for each control chart are shown in Table 4.4. The ARL

observed when estimating each of these control limits is also given. The standard error is approximately equal to 1 for all ARLs.

Table 4.4: Control limits for each control chart used to evaluate the out-of-control ARL performance of the wavelet-based disease surveillance method

			Control Chart				
Model	Sub-regions	Baseline Rate	Chi-Square CL (ARL)	MEWMA CL (ARL)	Weighted χ^2 CL (ARL)	One-Sided Subregion CL (ARL)	Two-Sided Subregion CL (ARL)
Canonical Link	16	1	31.77 (200.87)	13153 (200.62)		9.52 (200.01)	9.53 (199.59)
		10	34.08 (200.60)	41.32 (200.83)	2.73 (199.99)	6.63 (199.55)	6.80 (199.38)
		50	34.23 (199.92)	33.81 (200.05)	3.29 (200.49)	6.41 (200.78)	6.85 (200.45)
	32	1	47.57 (199.62)	21457 (201.46)		10.79 (200.96)	10.77 (200.10)
		10	55.49 (199.93)	68.89 (200.46)	2.93 (199.07)	7.36 (199.68)	7.51 (200.64)
		50	56.18 (199.18)	56.29 (200.54)	3.36 (199.22)	7.12 (199.51)	7.54 (199.25)
Identity Link	16	1	44.00 (≈206)	35.23 (199.85)	4.24 (199.57)	8.00 (≈185)	8.00 (≈185)
		10	35.64 (200.87)	33.28 (200.28)	3.54 (200.60)	6.75 (199.12)	6.99 (200.55)
		50	34.56 (200.57)	33.07 (201.37)	3.46 (199.95)	6.44 (199.51)	6.89 (200.67)
	32	1	67.00 (≈185)	57.44 (199.89)	4.58 (199.70)	8.99 (≈200)	8.99 (≈200)
		10	57.85 (200.67)	55.12 (200.01)	3.63 (199.61)	7.50 (199.51)	7.72 (199.34)
		50	56.65 (200.87)	54.88 (199.97)	3.51 (199.62)	7.15 (200.14)	7.58 (199.84)

When the canonical link model is used, there is no control limit for the Weighted χ^2 control chart when the baseline rate is 1 per 100,000 residents. In this case, there was not a control limit low enough that would result in an in-control ARL of 200. This is a result of the transformation performed on the weighted χ^2 statistic so that it would have an approximate standard normal distribution and could be monitored using a CUSUM control chart. The weighted χ^2 statistic, in its original form, is a weighted quadratic form and can not be less than zero; however, when the Wilson-Hilferty (1931) transformation is applied and the baseline incidence rate is very low, the resulting statistic can be negative on average, which is what occurred in this case. Although this issue does not occur when the identity link model is used, another issue presented itself for the chi-square, one-sided subregion, and two-sided subregion control charts when this model is used and the baseline rate is 1 per 100,000 resident. In this case, there was a jump in the in-control ARLs for continuous values of the control limit, and no control limit could be found to give an in-control ARL of 200. For these control charts, the control limit that produces an in-control ARL closest to 200 was used in the simulations.

4.2.2.1 Varying Region Size and Cluster Size

The out-of-control ARL performance was first evaluated for shifts in direction of the wavelet coefficient vector that reflected the formation of disease clusters of different size over the geographical region. These types of shifts were considered for regions that contain 16 and 32 subregions in the wavelet domain. The clusters formed by the shifts considered in the out-of-control ARL evaluation of the wavelet-based method are shown in Figures 4.3 and 4.4, for regions with 16 and 32 subregions, respectively. For both numbers of subregions, the out-of-control ARL performance was considered for clusters that cover the entire region, one half of the region, and one eighth of the region. The out-of-control ARL performance was also considered for clusters that cover one eighth of the region while there is a simultaneous decrease in the incidence rate, or negative cluster, that covers another eighth of the region. Several of these scenarios are similar to those shown in the demonstrations of the wavelet-based method in Section 3.4. These include the scenarios where a cluster forms over the entire region, where a cluster forms over one eighth of the region, and where a cluster forms over one eighth of the region while a negative cluster forms over another eighth of the region. These scenarios were used to evaluate the out-of-control ARL performance of the wavelet-based method for the same reasons they were selected for the demonstrations. In addition, the scenario where a cluster forms over one half of the region was evaluated to show the out-of-control ARL performance for a larger-sized cluster that does not cover the entire region.

To simulate the out-of-control scenarios shown in Figures 4.3 and 4.4, the shifts in the incidence rates from baseline were determined for each subregion included in a cluster so that the shift in the coefficient vector had a magnitude of approximately $\delta = 1$ based on the noncentrality parameter. These shifts were computed for the baseline incidence rates 1, 10, and 50 per 100,000 residents for the canonical and identity link models and are shown in Table 4.5. In the scenarios where there is a simultaneous increase in one eighth of the region and decrease in one eighth of the region, the increase and decrease from the baseline rate were kept equivalent unless this produced a negative incidence rate in the subregions where a decrease occurs. In this case, the incidence rates in the subregions included in the negative cluster were set equal to zero and the incidence rates in the subregions included in the cluster were determined so that the noncentrality parameter was approximately equal to 1. Then,

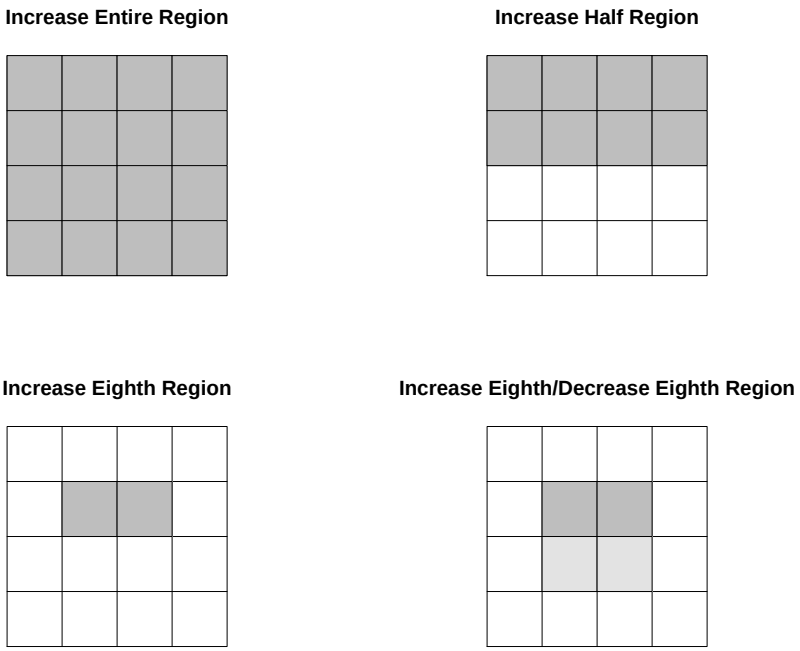


Figure 4.3: Cluster scenarios for out-of-control ARL simulations with 16 subregions

for each time interval, a random independent observation consisting of counts for each subregion was generated from the Poisson distributions corresponding to the specified baseline rate and the out-of-control rates. The population in each subregion remained at 100,000 for all time intervals and no covariate adjustments were made. After each observation was obtained, statistics to be plotted on each control chart were calculated for both the canonical and identity link models. These values were then compared to the appropriate control limits in Table 4.4. This process was repeated for each control chart until the chart signaled and one run length was produced. To estimate the out-of-control ARL, 10,000 run lengths were obtained under these same conditions for each out-of-control scenario and then averaged. The out-of-control ARL estimate computed for each scenario corresponds to the zero-state ARL. In this application, the zero-state ARL is the out-of-control ARL assuming the cluster formation is present from the start of monitoring.

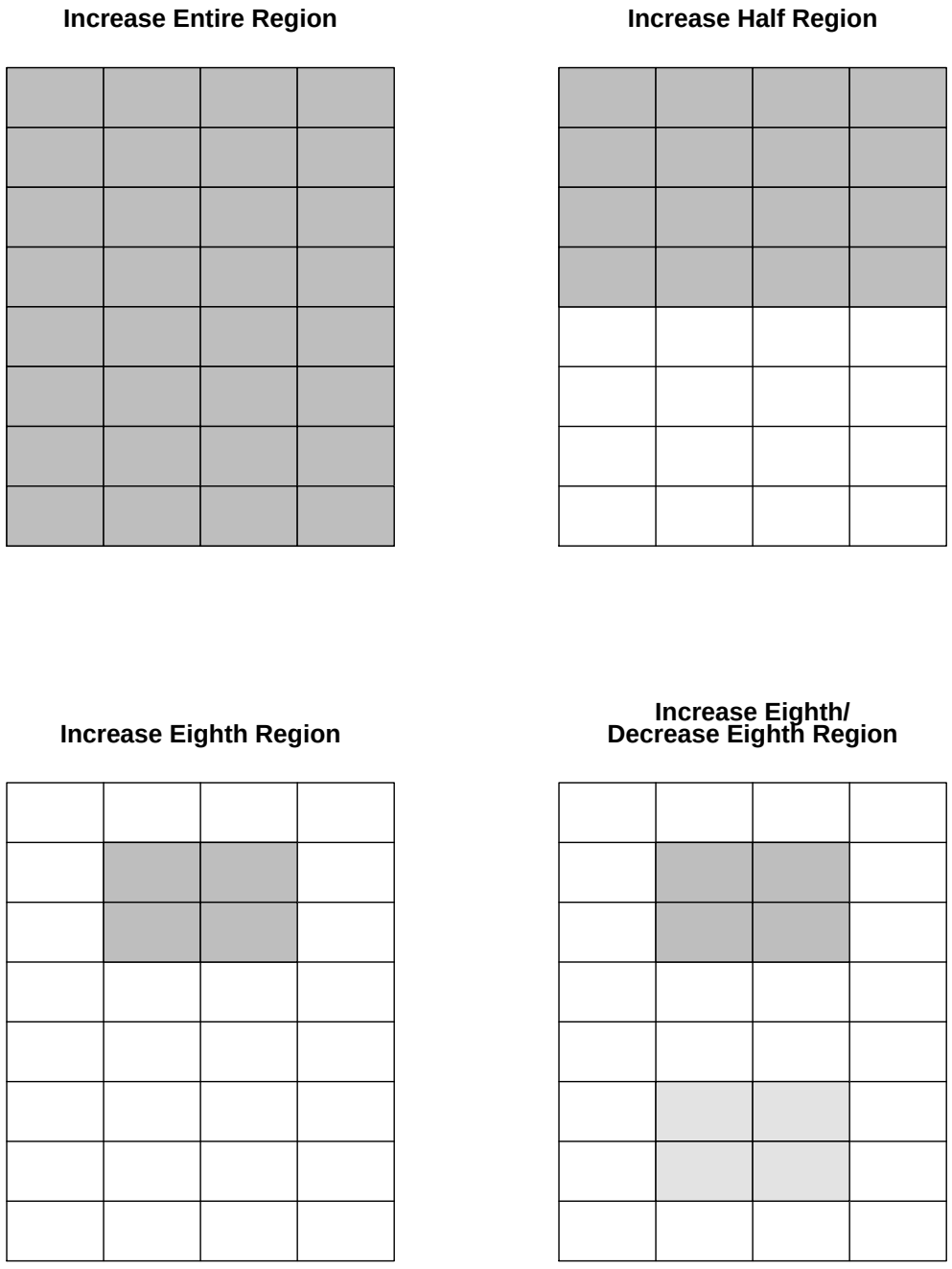


Figure 4.4: Cluster scenarios for out-of-control ARL simulations with 32 subregions

Table 4.5: Incidence rates per 100,000 residents for each out-of-control scenario representing changes in region and cluster size

			Baseline Rate (per 100,000 residents)		
			1	10	50
Model	Sub-regions	Out-of-Control Scenario	Decrease/ Increase	Decrease/ Increase	Decrease/ Increase
Canonical Link	16	Increase Over Entire Region	/1.355	/11.119	/52.500
		Increase Over Half of Region	/1.503	/11.583	/53.536
		Increase Over Eighth of Region	/2.021	/13.172	/57.076
		Decrease & Increase Over Eighth of Region	0.000/2.021	7.759/12.241	44.998/55.002
	32	Increase Over Entire Region	/1.251	/10.791	/51.768
		Increase Over Half of Region	/1.355	/11.119	/52.500
		Increase Over Eighth of Region	/1.716	/12.240	/55.002
		Decrease & Increase Over Eighth of Region	0.493/1.507	8.417/11.583	46.464/53.536
Identity Link	16	Increase Over Entire Region	/1.354	/11.118	/52.500
		Increase Over Half of Region	/1.500	/11.581	/53.536
		Increase Over Eighth of Region	/2.000	/13.162	/57.071
		Decrease & Increase Over Eighth of Region	0.293/1.707	7.764/12.236	45.000/55.000
	32	Increase Over Entire Region	/1.250	/10.791	/51.768
		Increase Over Half of Region	/1.354	/11.118	/52.500
		Increase Over Eighth of Region	/1.707	/12.236	/55.000
		Decrease & Increase Over Eighth of Region	0.500/1.500	8.419/11.581	46.464/53.536

Tables 4.6 and 4.7 contain the out-of-control ARL simulation results for the canonical and identity link models, respectively. To show the precision of the ARL estimates, the standard errors, SE, are also given. In order to identify the patterns in these out-of-control ARL results more easily, figures of the out-of-control ARL estimates are presented in Appendix E for each out-of-control scenario.

Discussion of Results Within Each Control Chart

For each control chart evaluated, the out-of-control ARLs are lower when there are 16 subregions as opposed to 32 subregions across all four of the out-of-control scenarios, regardless of the baseline count or model used, with two exceptions. When the baseline count is equal to 1 and the cluster covers the entire region, the chi-square control chart has comparable out-of-control performance when there are either 16 or 32 subregions, for both the canonical and identity link models. Also, the MEWMA control chart has larger out-of-control ARLs when there are 16 subregions rather than 32, when the baseline count is 1 and the canonical link model is used. Overall, this indicates that

Table 4.6: Out-of-control ARL estimates using the Poisson regression canonical link model for two different numbers of subregions, four different out-of-control scenarios, and three baseline rates per 100,000 residents

		Control Chart				
Subregions	Baseline Rate	Chi-Square	MEWMA	Weighted χ^2	One-Sided Subregion	Two-Sided Subregion
		Incidence Rate Increase Over the Entire Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	13.1 (0.13)	158.1 (1.45)		19.8 (0.07)	19.8 (0.07)
	10	25.4 (0.26)	28.0 (0.22)	9.0 (0.07)	16.1 (0.08)	16.9 (0.09)
	50	39.3 (0.39)	13.2 (0.08)	10.9 (0.08)	16.1 (0.08)	18.7 (0.10)
32	1	13.4 (0.13)	50.3 (0.37)		26.9 (0.10)	26.8 (0.09)
	10	32.0 (0.32)	82.0 (0.75)	19.3 (0.16)	22.4 (0.13)	23.7 (0.14)
	50	52.8 (0.52)	20.5 (0.14)	30.3 (0.26)	23.0 (0.13)	27.3 (0.16)
		Incidence Rate Increase Over One Half of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	20.1 (0.19)	70.6 (0.58)		17.7 (0.06)	17.7 (0.06)
	10	31.4 (0.32)	21.0 (0.15)	12.3 (0.10)	14.1 (0.07)	14.7 (0.07)
	50	44.4 (0.44)	12.5 (0.07)	14.4 (0.11)	14.2 (0.07)	15.8 (0.07)
32	1	21.8 (0.21)	31.0 (0.18)		24.4 (0.08)	24.3 (0.08)
	10	41.7 (0.41)	44.3 (0.36)	26.0 (0.23)	19.9 (0.11)	20.8 (0.11)
	50	60.3 (0.60)	18.1 (0.12)	36.0 (0.32)	20.1 (0.10)	22.7 (0.12)
		Incidence Rate Increase Over One Eighth of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	33.7 (0.33)	37.1 (0.25)		13.1 (0.05)	13.1 (0.05)
	10	41.3 (0.41)	15.9 (0.10)	30.4 (0.27)	9.9 (0.04)	10.1 (0.04)
	50	50.1 (0.50)	11.7 (0.06)	33.3 (0.30)	9.7 (0.04)	10.4 (0.04)
32	1	41.5 (0.41)	21.4 (0.09)		17.9 (0.06)	17.8 (0.06)
	10	56.8 (0.56)	26.3 (0.19)	38.1 (0.35)	13.7 (0.06)	14.1 (0.06)
	50	71.0 (0.70)	16.2 (0.10)	44.1 (0.41)	13.6 (0.06)	14.7 (0.06)
		Incidence Rate Increase & Decrease Over One Eighth of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	42.9 (0.43)	61.4 (0.48)		13.1 (0.05)	13.1 (0.05)
	10	69.3 (0.69)	13.8 (0.08)	54.0 (0.50)	16.5 (0.09)	13.9 (0.06)
	50	65.5 (0.65)	11.3 (0.06)	45.1 (0.42)	16.2 (0.09)	13.4 (0.05)
32	1	105.8 (1.07)	17.6 (0.06)		26.7 (0.11)	26.6 (0.11)
	10	95.3 (0.94)	20.1 (0.12)	64.0 (0.61)	23.6 (0.14)	20.5 (0.10)
	50	90.7 (0.90)	15.2 (0.09)	56.2 (0.53)	23.8 (0.14)	19.9 (0.09)

Table 4.7: Out-of-control ARL estimates using the Poisson regression identity link model for two different numbers of subregions, four different out-of-control scenarios, and three baseline rates per 100,000 residents

		Control Chart				
Subregions	Baseline Rate	Chi-Square	MEWMA	Weighted χ^2	One-Sided Subregion	Two-Sided Subregion
		Incidence Rate Increase Over the Entire Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	18.1 (0.17)	8.6 (0.05)	6.6 (0.04)	15.5 (0.08)	15.5 (0.08)
	10	28.0 (0.27)	9.9 (0.05)	8.9 (0.06)	15.9 (0.08)	17.3 (0.09)
	50	39.8 (0.39)	10.4 (0.05)	10.4 (0.08)	16.1 (0.08)	18.8 (0.10)
32	1	18.4 (0.18)	10.0 (0.05)	8.9 (0.06)	21.1 (0.12)	21.1 (0.12)
	10	35.2 (0.35)	12.5 (0.07)	18.9 (0.15)	22.3 (0.13)	24.3 (0.14)
	50	53.3 (0.52)	13.6 (0.08)	29.9 (0.26)	22.9 (0.13)	27.1 (0.16)
		Incidence Rate Increase Over One Half of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	25.1 (0.24)	9.6 (0.05)	9.2 (0.06)	13.8 (0.07)	13.8 (0.07)
	10	33.4 (0.33)	10.3 (0.05)	11.8 (0.09)	14.0 (0.07)	14.9 (0.07)
	50	44.2 (0.44)	10.7 (0.05)	13.8 (0.11)	14.1 (0.07)	15.8 (0.07)
32	1	26.7 (0.26)	11.3 (0.06)	12.6 (0.09)	19.0 (0.10)	19.0 (0.10)
	10	44.1 (0.44)	13.0 (0.07)	24.9 (0.21)	19.7 (0.11)	21.2 (0.11)
	50	61.7 (0.60)	13.8 (0.08)	35.3 (0.31)	20.2 (0.10)	22.9 (0.12)
		Incidence Rate Increase Over One Eighth of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	36.7 (0.37)	10.8 (0.06)	19.7 (0.17)	10.5 (0.05)	10.5 (0.05)
	10	43.5 (0.43)	10.8 (0.06)	25.9 (0.22)	9.9 (0.04)	10.3 (0.04)
	50	50.6 (0.50)	11.1 (0.06)	31.0 (0.27)	9.7 (0.04)	10.5 (0.04)
32	1	43.5 (0.43)	13.3 (0.08)	23.5 (0.20)	14.2 (0.07)	14.2 (0.07)
	10	59.0 (0.58)	13.7 (0.08)	35.4 (0.32)	13.6 (0.06)	14.2 (0.06)
	50	71.9 (0.70)	14.4 (0.08)	43.3 (0.39)	13.6 (0.06)	14.7 (0.06)
		Incidence Rate Increase & Decrease Over One Eighth of the Region				
		ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)	ARL (SE)
16	1	75.7 (0.75)	12.7 (0.07)	54.1 (0.51)	17.3 (0.10)	15.9 (0.08)
	10	69.2 (0.69)	11.4 (0.06)	45.4 (0.42)	16.5 (0.09)	13.7 (0.06)
	50	65.1 (0.65)	11.1 (0.06)	44.0 (0.41)	16.2 (0.09)	13.5 (0.06)
32	1	93.3 (0.93)	16.7 (0.10)	62.9 (0.58)	24.2 (0.15)	23.3 (0.13)
	10	92.3 (0.91)	14.9 (0.08)	57.4 (0.54)	23.5 (0.14)	20.3 (0.10)
	50	90.7 (0.90)	14.9 (0.09)	55.8 (0.52)	23.9 (0.14)	19.9 (0.09)

the control charts have better performance for all out-of-control scenarios when the number of subregions is lower. This is expected because there is a reduction in the dimensions of the parameter space when the number of subregions is lower.

Another characteristic of out-of-control ARL performance seen in these results is that the model with the best performance changes depending on the baseline incidence count for a majority of the control charts. At low baseline counts, the chi-square control chart has lower out-of-control ARLs when the canonical link model is used, but as the baseline count increases the ARLs for both the canonical and identity link models become comparable. For the Weighted χ^2 control chart, the performance is better for the baseline count of 10 compared to 50 when the identity link model is used for the out-of-control scenarios where a cluster covers the whole, one half, or one eighth of the region. In all other cases, the performance for each model is comparable for baseline counts of 10 and 50 for this control chart. When the baseline count is equal to 1, the one-sided and two-sided subregion control charts have lower out-of-control ARLs when the identity link model is used. The out-of-control ARL performance for these control charts is comparable for the canonical and identity link models when the baseline counts are 10 or 50, except in the out-of-control scenario where there is both a cluster and negative cluster present. The MEWMA control chart is the only chart that has better out-of-control ARL performance for the identity link model over the canonical link model regardless of the baseline count in all out-of-control scenarios. For this control chart, the out-of-control ARLs for the canonical link model begin to approach the ARLs of the identity link model as the baseline count increases, but they are still higher when compared to the ARLs of the identity link model when the baseline count reaches 50.

The out-of-control ARL performance for each control chart also changes depending on the out-of-control scenario considered, and with the MEWMA control chart it also depends on the model used. When the MEWMA control chart is used with the canonical link model, the performance of the control chart improves as the cluster size decreases from the entire region to one eighth of the region. This chart also has good performance, relative to the other out-of-control scenarios, when there is a simultaneous cluster and negative cluster forming in the region and the canonical link model is used. When the identity link model is used, the MEWMA control chart has

slightly better out-of-control ARL performance for larger clusters and the performance when there is both a cluster and negative cluster is slightly worse relative to the other out-of-control scenarios. The chi-square and Weighted χ^2 control charts have the best out-of-control performance when the cluster covers the entire region. For both of these charts, as the cluster size decreases, the out-of-control ARLs increase showing a decrease in performance. The chi-square and Weighted χ^2 control charts perform at their worst for the scenario where there is both a cluster and negative cluster present, when compared to the other out-of-control scenarios. The out-of-control ARL performance for the one-sided and two-sided subregion control charts improves as the cluster size decreases, regardless of the model used. In the scenario where there is a simultaneous cluster and negative cluster, the out-of-control ARLs are comparable to the ARLs for some of the other out-of-control scenarios, but these one-sided and two-sided charts do not perform at their best in this situation.

Discussion of Results Across Control Charts

When comparing the out-of-control ARL performance of the control charts to one another, no control chart is best in its ability to detect clusters of disease in all of the out-of-control scenarios. In many cases, the control chart that performs best overall for each out-of-control scenario depends on the model used, the baseline count, and the number of subregions. For the canonical link model, the chi-square control chart is best for detecting clusters covering the entire or half of the region when the baseline count is 1. When the baseline count increases to 10 in these scenarios, the Weighted χ^2 control chart has the best out-of-control ARL performance, with the exception of when the cluster covers half of the region and the number of subregions is equal to 32. In this case, the one-sided and two-sided subregions control charts have the best out-of-control ARL performance for the canonical link model. At the baseline count of 50, the Weighted χ^2 control chart is also the best performer when the cluster covers the entire region and the number of subregions is 16. In all other cases when the baseline count is 50 and the canonical link model is used, the MEWMA control chart has the best out-of-control ARL performance. In the scenario where a cluster forms over one eighth of the region and the canonical link model is used, the one-sided and two-sided subregion control charts have the best out-of-control ARL performance, regardless of the baseline

count and number of subregions. When there is a cluster and negative cluster that each cover one eighth of the region, the one-sided and two-sided control charts have the best out-of-control ARL performance for 16 subregions and the MEWMA control chart has the best performance for 32 subregions, when the baseline count is 1 and the canonical link model is used. In this same out-of-control scenario, the MEWMA and two-sided subregion control charts have the best performance when the baseline count is 10 and the MEWMA control chart has the best performance when the baseline count is 50. Overall, across all four out-of-control scenarios, the MEWMA control chart and the subregion control charts perform well and the chi-square control chart has the worst performance when the canonical link model is used.

When the identity link model is used, the pattern of out-of-control ARL performance for the control charts changes, and the control charts that perform best overall for each out-of-control scenario differ from those that perform best when the canonical link model is used. In the out-of-control scenario where a cluster covers the entire region, the Weighted χ^2 or MEWMA control chart has the best out-of-control ARL performance when the identity link model is used depending on the baseline count and number of subregions. In this scenario, the Weighted χ^2 control chart has the best performance for the subregion and baseline count combinations 16 and 1, 16 and 10, and 32 and 1, and the MEWMA control chart has the best performance for the subregion and baseline count combinations 32 and 10 and 32 and 50. When there are 16 subregions and the baseline count is 50, the Weighted χ^2 and MEWMA control charts have comparable performance for this scenario. When there is a cluster that covers half of the region and the identity link model is used, the MEWMA control chart has the best in-control ARL performance in all cases when compared to the other control charts, except when there are 16 subregions and the baseline count is 1. In this case, the MEWMA and Weighted χ^2 charts have comparable out-of-control performance that is better than the other control charts. In the out-of-control scenario, where there is a cluster that covers one eighth of the region, the MEWMA, one-sided subregion, and two-sided subregion control charts all have comparable out-of-control ARL performance when the identity link model is used that is better than the Weighted χ^2 and chi-square control charts, regardless of the baseline count or number of subregions. In the scenario where there is a cluster and negative cluster present, the MEWMA control

chart has the best out-of-control ARL performance for each baseline count and number of subregions combination when the identity link model is used. Overall, across all four out-of-control scenarios when the identity link model is used, the MEWMA control chart performs best and the chi-square control chart performs worst.

When the out-of-control ARL performance of the control charts that perform best in each out-of-control scenario is compared for the two models, the use of the identity link model leads to control charts with a lower or comparable out-of-control ARL when compared to the control charts with the best performance using the canonical link model in all cases. When the identity link model is used, the MEWMA control chart performs best in a majority of the out-of-control scenarios, and in the cases where it is not the best performer, it has an out-of-control ARL almost as low as the control chart with the lowest out-of-control ARL. Therefore, if one were to choose a control chart and model that would perform well in most out-of-control scenarios, the MEWMA control chart using the identity link function would be best.

4.2.2.2 Varying Cluster Shape and Location

In addition to evaluating the out-of-control ARL performance of the control charts in the wavelet-based disease surveillance method for clusters of different size, it is also important to determine the impact that clusters of different shape and location have on the out-of-control ARL performance. The performance of the Weighted χ^2 control chart is of particular interest when assessing the impact of changes in cluster shape and location. This is because the statistic used in the Weighted χ^2 control chart weights the contribution of the wavelet coefficients differently depending on their corresponding resolution. Since each wavelet coefficient is associated with a specific area of the region, the out-of-control ARL performance of the Weighted χ^2 control chart will differ depending on the location of a cluster within the geographical region. Since the shape of the cluster is related to location within the wavelet domain, this will also impact the out-of-control ARL performance. The other control charts that can be used with the wavelet-based method will have the same out-of-control ARL performance for clusters in different locations of the region and for clusters of differing shape because the wavelet coefficients are weighted equally in the statistics monitored by these charts.

In order to investigate the impact of cluster shape and location on the Weighted χ^2 control chart, several out-of-control scenarios were considered. In each scenario the region and cluster size remained constant so that only the impact of cluster shape and location could be examined. The region size selected was 16, meaning there were 16 subregions in the wavelet domain for each region, and each cluster covered 4 subregions, which equates to one quarter of the region. The 12 out-of-control scenarios considered are shown in Figure 4.5, where the shaded areas in the wavelet domain indicate the subregions with increased incidence rates that form disease clusters. The shapes of these cluster formations were selected because they represent every possible cluster formation covering one quarter of the region, where all of the subregions included in the cluster are adjacent to one another. The locations for each different shape were selected because the out-of-control ARL performance for all possible cluster locations is equivalent to the out-of-control ARL performance of one of the 12 scenarios evaluated. For each scenario evaluated, the region can be rotated or translated to change the cluster location, but this has no impact on the out-of-control ARL performance.

To simulate the 12 out-of-control scenarios represented in Figure 4.5, shifts in the incidence rates from baseline were computed for the subregions that make up each cluster so that the shift in the coefficient vector gave a noncentrality parameter of $\delta = 1$. These shifts were computed for the baseline incidence rates 1, 10, and 50 per 100,000 residents for the canonical and identity link models and are shown in Table 4.8. Then, the out-of-control ARLs were computed for each control chart in the same way they were for the evaluation based on varying region and cluster size in Section 4.2.2.1, where a fixed population size of 100,000 was assumed over time and no covariate adjustments were made. As with the out-of-control ARLs estimated for investigating control chart performance for differing region and cluster sizes, the out-of-control ARL estimate computed for each cluster shape and location scenario corresponds to the zero-state ARL.

Tables 4.9 and 4.10 contain the out-of-control ARL simulation results for the Weighted χ^2 control charts in each of the 12 out-of-control scenarios using the the canonical and identity link models, respectively. Plots of the out-of-control ARL results for the Weighted χ^2 control chart are also provided in Appendix E. The out-of-control ARL results for other control charts that can be used in the wavelet-based

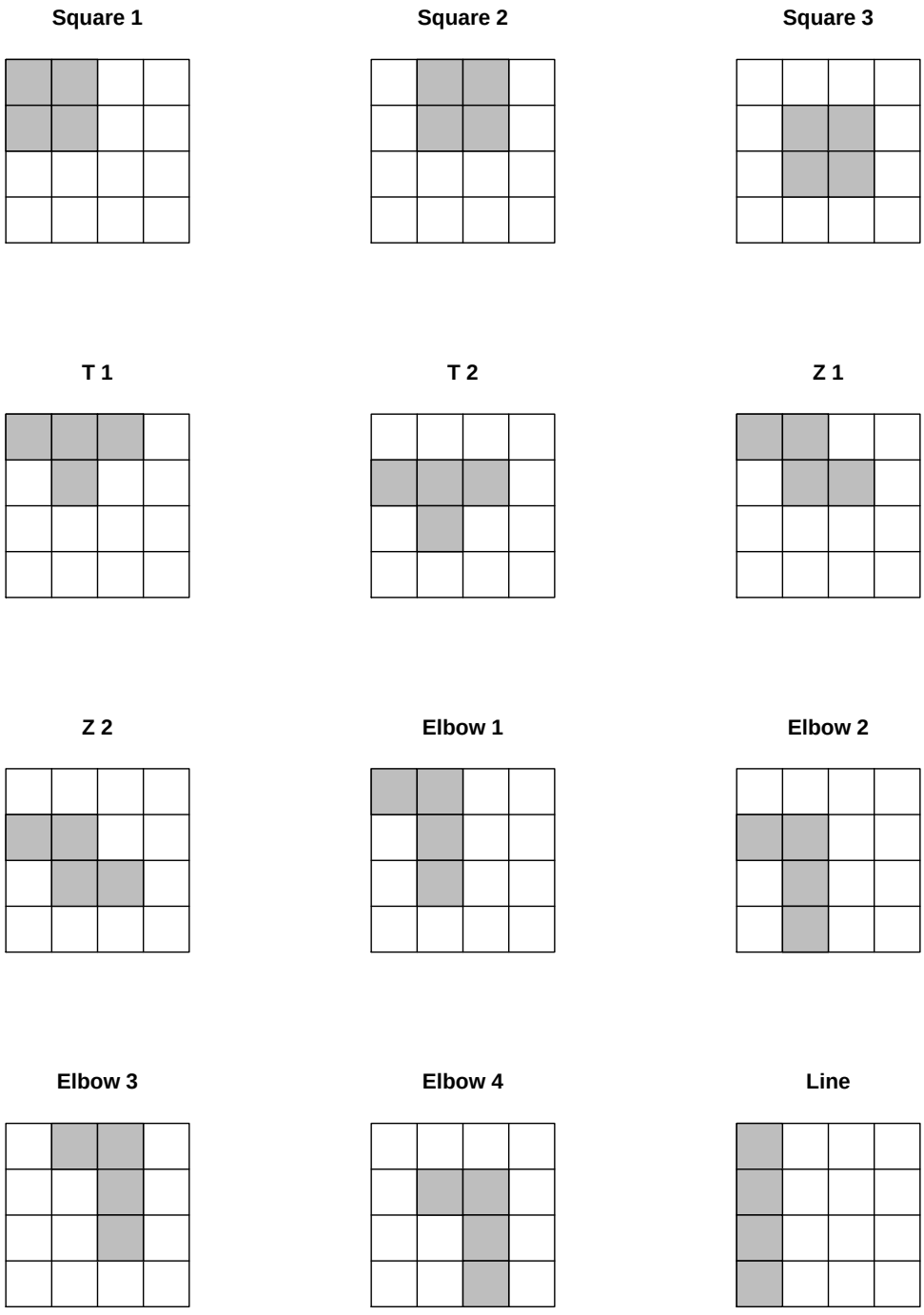


Figure 4.5: Cluster scenarios for out-of-control ARL simulations with differing cluster shapes and locations

Table 4.8: Incidence rates per 100,000 residents for each out-of-control scenario representing changes in cluster shape and location

	Baseline Rate (per 100,000 residents)		
Model	1	10	50
Canonical Link	1.716	12.240	55.002
Identity Link	1.707	12.236	55.000

method are shown in each table for comparison, even though they are not expected to change for these out-of-control scenarios. The standard errors, SE, are provided in the tables to show the precision of the ARL estimates. These results indicate that when the Weighted χ^2 control chart is used, the cluster shapes and locations that lead to the best out-of-control ARL performance are those where the low resolution wavelet coefficients in the Poisson regression model are large in magnitude. When there are 16 subregions in the wavelet domain, the lowest resolution wavelet functions split the wavelet domain in half in each dimension creating four quadrants in the corners of the wavelet domain. Generally, the cluster shapes and locations that cover fewer quadrants can be detected more quickly by the Weighted χ^2 control chart because in these cases the low resolution wavelet coefficients are larger in magnitude. An extreme example of this phenomenon is shown with the shape and location combinations represented by Square 1 and Square 3. The Square 1 cluster is completely contained within one quadrant, and this cluster is detected more quickly for all baseline counts and both models. On the other hand, the Square 3 cluster covers all four quadrants, and this cluster is the most difficult to detect regardless of the baseline count or model. When comparing the out-of-control ARL performance for the Weighted χ^2 control chart overall in these out-of-control scenarios, the ARLs are slightly lower when using the identity link model than when the canonical link model is used. Also, the out-of-control ARLs increase as the baseline count increases for each shape, regardless of the model used.

When the performance of the Weighted χ^2 control chart is compared to the other control charts that can be used in the wavelet-based method, it generally has poor performance. The chi-square, MEWMA, one-sided subregion, and two-sided subregion control charts have the same performance for each cluster shape and location combi-

Table 4.9: Out-of-control ARL estimates using the Poisson regression canonical link model for 12 different cluster shape and location scenarios and three baseline rates per 100,000 residents

Cluster Shape and Location	Baseline Rate	Control Chart				
		Chi-Square	MEWMA	Weighted χ^2	One-Sided Subregion	Two-Sided Subregion
		ARL (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)
Square 1	1	26.8 (0.26)	46.8 (0.34)		15.2 (0.06)	15.2 (0.06)
	10	37.7 (0.37)	17.4 (0.11)	16.4 (0.14)	11.8 (0.05)	12.2 (0.05)
	50	48.9 (0.48)	11.9 (0.06)	18.5 (0.15)	11.7 (0.05)	12.8 (0.05)
Square 2	1	26.5 (0.26)	46.2 (0.33)		15.3 (0.06)	15.3 (0.06)
	10	37.0 (0.37)	17.5 (0.11)	20.2 (0.18)	11.9 (0.05)	12.3 (0.06)
	50	48.8 (0.48)	11.9 (0.06)	22.7 (0.20)	11.7 (0.05)	12.7 (0.05)
Square 3	1	26.9 (0.26)	46.7 (0.34)		15.3 (0.06)	15.3 (0.06)
	10	37.7 (0.38)	17.5 (0.11)	23.8 (0.22)	12.0 (0.05)	12.4 (0.06)
	50	47.5 (0.47)	11.8 (0.06)	26.8 (0.24)	11.7 (0.05)	12.8 (0.06)
T 1	1	27.1 (0.27)	46.2 (0.33)		15.3 (0.06)	15.3 (0.06)
	10	37.1 (0.37)	17.8 (0.12)	19.5 (0.17)	11.8 (0.05)	12.2 (0.05)
	50	48.5 (0.48)	12.0 (0.07)	22.0 (0.19)	11.8 (0.05)	12.9 (0.06)
T 2	1	26.8 (0.26)	45.7 (0.32)		15.2 (0.06)	15.3 (0.06)
	10	37.3 (0.36)	17.7 (0.11)	21.9 (0.19)	11.9 (0.05)	12.3 (0.05)
	50	48.8 (0.49)	12.0 (0.07)	24.9 (0.22)	11.8 (0.05)	12.9 (0.05)
Z 1	1	27.0 (0.27)	47.1 (0.34)		15.3 (0.06)	15.3 (0.06)
	10	37.5 (0.37)	17.6 (0.11)	19.7 (0.17)	11.8 (0.05)	12.2 (0.05)
	50	48.6 (0.48)	11.9 (0.06)	22.2 (0.19)	11.8 (0.05)	12.9 (0.05)
Z 2	1	26.7 (0.26)	46.6 (0.33)		15.3 (0.06)	15.3 (0.06)
	10	37.1 (0.37)	17.5 (0.11)	23.1 (0.21)	11.9 (0.05)	12.3 (0.05)
	50	48.6 (0.48)	12.0 (0.07)	25.5 (0.22)	11.8 (0.05)	12.8 (0.05)
Elbow 1	1	26.9 (0.27)	46.9 (0.34)		15.3 (0.06)	15.3 (0.06)
	10	38.1 (0.38)	17.6 (0.11)	19.4 (0.17)	11.9 (0.05)	12.3 (0.05)
	50	48.6 (0.48)	12.0 (0.06)	22.1 (0.19)	11.8 (0.05)	12.9 (0.05)
Elbow 2	1	27.4 (0.27)	46.4 (0.34)		15.3 (0.06)	15.3 (0.06)
	10	37.0 (0.36)	17.6 (0.11)	20.0 (0.18)	11.9 (0.05)	12.3 (0.06)
	50	49.0 (0.48)	11.9 (0.07)	22.8 (0.19)	11.8 (0.05)	12.9 (0.05)
Elbow 3	1	27.0 (0.27)	46.2 (0.33)		15.3 (0.06)	15.3 (0.06)
	10	37.4 (0.37)	17.7 (0.11)	21.7 (0.19)	11.9 (0.05)	12.3 (0.06)
	50	49.0 (0.49)	12.0 (0.07)	25.0 (0.22)	11.7 (0.05)	12.8 (0.05)
Elbow 4	1	26.9 (0.26)	47.2 (0.34)		15.3 (0.06)	15.3 (0.06)
	10	37.4 (0.37)	17.7 (0.12)	22.2 (0.20)	11.8 (0.05)	12.2 (0.05)
	50	48.8 (0.48)	12.0 (0.07)	24.8 (0.22)	11.7 (0.05)	12.9 (0.05)
Line	1	26.9 (0.26)	46.2 (0.33)		15.3 (0.06)	15.3 (0.06)
	10	37.5 (0.37)	17.6 (0.11)	19.5 (0.17)	11.9 (0.05)	12.3 (0.06)
	50	47.8 (0.48)	12.0 (0.07)	21.8 (0.19)	11.8 (0.05)	12.9 (0.05)

Table 4.10: Out-of-control ARL estimates using the Poisson regression identity link model for 12 different cluster shape and location scenarios and three baseline rates per 100,000 residents

Cluster Shape and Location	Baseline Rate	Control Chart				
		Chi-Square	MEWMA	Weighted χ^2	One-Sided Subregion	Two-Sided Subregion
		ARL (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)	*ARL* (*SE*)
Square 1	1	31.4 (0.31)	10.3 (0.06)	12.3 (0.09)	12.1 (0.06)	12.1 (0.06)
	10	38.8 (0.38)	10.6 (0.06)	15.3 (0.12)	11.9 (0.05)	12.5 (0.05)
	50	49.7 (0.48)	10.8 (0.06)	17.5 (0.14)	11.7 (0.05)	12.9 (0.06)
Square 2	1	31.3 (0.30)	10.3 (0.06)	13.9 (0.11)	12.2 (0.06)	12.2 (0.06)
	10	39.1 (0.39)	10.5 (0.06)	17.8 (0.15)	11.8 (0.05)	12.5 (0.06)
	50	49.8 (0.49)	10.8 (0.06)	21.3 (0.18)	11.7 (0.05)	12.8 (0.05)
Square 3	1	31.5 (0.31)	10.2 (0.06)	14.7 (0.12)	12.3 (0.06)	12.3 (0.06)
	10	39.2 (0.39)	10.6 (0.06)	20.7 (0.18)	11.8 (0.05)	12.4 (0.05)
	50	47.9 (0.48)	10.8 (0.06)	24.7 (0.21)	11.7 (0.05)	12.9 (0.05)
T 1	1	31.4 (0.31)	10.4 (0.06)	13.7 (0.11)	12.2 (0.06)	12.2 (0.06)
	10	39.1 (0.38)	10.6 (0.06)	17.4 (0.14)	11.9 (0.05)	12.5 (0.06)
	50	48.9 (0.49)	10.8 (0.06)	20.5 (0.17)	11.7 (0.05)	12.9 (0.05)
T 2	1	31.4 (0.31)	10.2 (0.06)	14.1 (0.11)	12.2 (0.06)	12.2 (0.06)
	10	39.4 (0.39)	10.6 (0.06)	19.3 (0.16)	11.8 (0.05)	12.5 (0.05)
	50	49.2 (0.49)	10.8 (0.06)	23.1 (0.19)	11.8 (0.05)	12.9 (0.05)
Z 1	1	31.2 (0.30)	10.3 (0.06)	13.6 (0.11)	12.1 (0.06)	12.1 (0.06)
	10	39.0 (0.39)	10.5 (0.06)	18.0 (0.15)	11.9 (0.05)	12.5 (0.05)
	50	49.1 (0.48)	10.8 (0.06)	20.6 (0.17)	11.8 (0.05)	12.9 (0.05)
Z 2	1	31.6 (0.31)	10.2 (0.06)	14.5 (0.12)	12.3 (0.06)	12.3 (0.06)
	10	39.6 (0.40)	10.5 (0.06)	19.8 (0.17)	11.7 (0.05)	12.4 (0.05)
	50	48.8 (0.49)	10.8 (0.06)	23.5 (0.20)	11.8 (0.05)	13.0 (0.06)
Elbow 1	1	31.2 (0.30)	10.2 (0.06)	13.3 (0.10)	12.2 (0.06)	12.2 (0.06)
	10	39.2 (0.39)	10.5 (0.06)	17.6 (0.15)	11.8 (0.05)	12.5 (0.05)
	50	49.2 (0.49)	10.8 (0.05)	20.7 (0.17)	11.8 (0.05)	12.9 (0.05)
Elbow 2	1	32.3 (0.32)	10.3 (0.06)	13.8 (0.11)	12.2 (0.06)	12.2 (0.06)
	10	39.9 (0.39)	10.7 (0.06)	18.1 (0.15)	11.9 (0.05)	12.6 (0.06)
	50	49.6 (0.49)	10.8 (0.06)	21.4 (0.18)	11.8 (0.05)	12.9 (0.06)
Elbow 3	1	31.9 (0.32)	10.3 (0.06)	14.5 (0.11)	12.4 (0.06)	12.4 (0.06)
	10	39.2 (0.38)	10.5 (0.06)	19.4 (0.16)	11.9 (0.05)	12.5 (0.06)
	50	49.4 (0.49)	10.9 (0.06)	23.2 (0.20)	11.7 (0.05)	12.9 (0.05)
Elbow 4	1	31.0 (0.30)	10.3 (0.06)	14.4 (0.11)	12.4 (0.06)	12.4 (0.06)
	10	39.4 (0.38)	10.6 (0.06)	19.4 (0.17)	11.8 (0.05)	12.5 (0.05)
	50	49.1 (0.49)	10.8 (0.06)	23.0 (0.20)	11.8 (0.05)	12.9 (0.05)
Line	1	31.1 (0.31)	10.3 (0.06)	13.3 (0.10)	12.2 (0.06)	12.2 (0.06)
	10	39.6 (0.40)	10.6 (0.06)	17.5 (0.14)	11.8 (0.05)	12.4 (0.06)
	50	48.5 (0.49)	10.9 (0.06)	20.5 (0.17)	11.8 (0.05)	13.0 (0.06)

nation as expected. For the identity link model, the MEWMA control chart has the lowest out-of-control ARLs and the one-sided and two-sided subregion control charts have out-of-control ARLs slightly above the values for the MEWMA control chart, for all of the baseline counts. All three of these control charts have better out-of-control ARL performance overall when compared to the Weighted χ^2 control chart using the identity link model for each out-of-control scenario. For the canonical link model, the MEWMA, one-sided subregion, and two-sided subregion control charts all generally have comparable out-of-control ARLs that are lower than the out-of-control ARLs for the Weighted χ^2 control chart, in each out-of-control scenario. These results are similar to those found in Section 4.2.2.1, where the out-of-control ARL performance of these control charts was compared for the detection of clusters of different size.

4.2.2.3 Performance of the Weighted χ^2 Control Chart

In the simulations performed to evaluate the out-of-control ARL performance of the wavelet-based method, the Weighted χ^2 control chart did not perform as well as anticipated compared to the MEWMA control chart as the cluster size decreased. Based on the weighting scheme used for the weighted χ^2 statistic, which was discussed in Section 3.3.3.3, the Weighted χ^2 chart is expected to perform well when the cluster sizes are large because the lower resolution wavelet coefficients are weighted more heavily. The control chart generally performed well in these scenarios, as expected, and outperformed the MEWMA control chart as well as the chi-square control chart in several cases. There was also an expectation, however, that the weighting scheme used for the weighted χ^2 statistic would lead to a control chart with better performance when the cluster size is small. This was suspected because the formation of small clusters causes a shift in both the high resolution and low resolution wavelet coefficients. As the cluster size decreased in the simulations, however, the MEWMA control chart outperformed the Weighted χ^2 control chart in terms of out-of-control ARL performance in all cases. Overall, the Weighted χ^2 control chart only performed better than the chi-square control chart, out of all control charts evaluated, when the cluster size decreased.

There was another issue that also became apparent when evaluating the Weighted χ^2 control chart. The distribution used to approximate the statistic monitored by this

chart does not provide a good approximation when the canonical link model is used and the baseline count is low. Due to the poor approximation, an appropriate control limit could not be found for monitoring purposes, which showed that this control chart has little use when the baseline counts are low. When the baseline counts are low, only the identity link model can be used in conjunction with the Weighted χ^2 control chart.

Although the Weighted χ^2 control chart did not perform as well as expected in terms of out-of-control ARL performance, the concept of monitoring a statistic that applies different weights to wavelet coefficients of different resolution may still be worthwhile for detecting clusters of varying size. Spitzner and Marshall (2008) developed an alternative wavelet-based monitoring method that uses a different weighted statistic to monitor clusters of disease. The weighted statistic used in this method is a weighted sum of the sequential sum-of-squares estimated using a wavelet-based Poisson regression model, where the weight for each sequential sum-of-squares, $SS(X_j|X_1, X_2, \ldots, X_{j-1})$, is $j^{-1/2}$ for regressors 1 through p. This alternative method differs from the wavelet-based method developed in this chapter in that the wavelet decomposition is applied directly to an irregularly-shaped region, rather than through a mapping to the wavelet domain. Although the weighted statistic used in the alternative method has only been evaluated for retrospective cluster detection thus far, the power of this statistic to detect clusters of many shapes and sizes retrospectively has proven to be high when compared to other statistics developed for this purpose. A similar weighted statistic could be developed for monitoring disease clusters in the wavelet domain for use in the wavelet-based method discussed in this chapter, by applying weights to the sequential sum-of-squares of the Poisson regression model for groups of regressors corresponding to the same resolution. A control chart used to monitor this weighted statistic may have better ability to detect clusters of varying size and shape than the control charts considered in this evaluation, even in cases where the baseline count is low.

It should also be noted that the Weighted χ^2 control chart is still useful in cases when one is interested in detecting clusters of a specific size. This application of the control chart is also discussed in Section 3.3.3.3 and requires applying a weighting scheme that weighs the higher or lower resolution wavelet coefficients more heavily depending on the size of the cluster one is interested in detecting. The weighting scheme used for the Weighted χ^2 control chart in this evaluation is one that puts more emphasis on

the lower resolution wavelet coefficients, and represents a possible weighting scheme if one is interested in detecting large clusters that cover a majority of the region. In this case, the Weighted χ^2 control chart performed well.

4.3 Female Respiratory Lung Cancer Case Study

In order to evaluate the wavelet-based disease surveillance method in a real-world scenario, this method was applied to the data on female respiratory lung cancer incidences in New Mexico collected through the SEER program. These data have been used previously by Kulldorff and Hjalmars (1999) to evaluate the population shift bias for tests of space-time interaction. As discussed in Section 3.1, these data consist of monthly female lung cancer incidences in each county of New Mexico from 1973 to 1991. There were a total of 228 observations over this period, which allowed for a portion of these data to be used to estimate the baseline incidence rate. The yearly estimated female population counts were also provided with the incidence counts, which allowed for population information to be used in the method. Additionally, the incidences were reported by age group, which would ideally be used as a covariate; however, the incidence counts were very low in these counties and further division of the data would not allow for stable baseline estimates to be obtained.

To estimate the baseline incidence rates of female lung cancer in each New Mexico county, the first quarter of the data were used as a historical data set, which consisted of observations from January of 1973 through September of 1977. To determine the baseline estimates, first, the incidence rates for each observation in the first quarter of the data were determined by dividing the count for a given month by the population estimate in the corresponding year in each county. After determining the monthly incidence rates for each county from these data, extreme rates in each county were identified and removed. To identify these outliers, the inter-quartile range (IQR) rule was applied. In each county, the IQR was determined for the first 57 observations, along with the first and third quartiles, denoted by Q_1 and Q_3, respectively. If an observation fell below $Q_1(i) - 1.5[IQR(i)]$ or above $Q_3(i) + 1.5[IQR(i)]$, for each subregion i, it was considered an outlier. In some cases there were so few non-zero rates, that that Q_1 and Q_3 were equal to zero, leading to an IQR of zero. In counties where

this occurred, the IQR rule was not applied and no outliers were identified. Once the outliers were removed from the first 57 observations, the baseline incidence rates were estimated for each county by averaging the monthly incidence rates for the remaining observations in the first quarter of the data. The baseline incidence rate estimates across the counties ranged from 0.129 per 100,000 females in McKinley county to 4.41 per 100,000 females in Catron county. The average female population in the counties over this time period ranged from 591 in Harding county to 190,314 in Bernalillo county. The city of Albuquerque is in Bernalillo county, which is the reason this county has the largest population of females. When comparing the estimated baseline rates and average female population of each county over the first quarter of the data, there was no obvious correlation between these two variables.

The estimated baseline rates in each county were low, indicating that the diagnosis of female respiratory lung cancer is a rare event. Based on the population sizes in these counties over the monitoring period, the incidence counts from October of 1977 to December of 1991 are expected to be low if the incidence rates remain close to the baseline estimates. In some cases, the expected count will be less than one per 100,000 females. In this type of scenario, the control charts using the identity link model performed best based on the out-of-control ARL results presented in Section 4.2.2. Therefore, this model was used in the analysis of the female lung cancer data. The chi-square, MEWMA, and Weighted χ^2 control charts were all used to monitor these data from October 1977 to December 1991 to compare their ability to detect clusters of disease in this application. Based on the results shown in Section 4.2.2, the MEWMA and Weighted χ^2 control charts are expected to perform best in this scenario.

In addition to estimating the baseline incidence rates for each county, a reasonable mapping of the subregions to the wavelet domain and the control limits for the charts used in this application were determined prior to monitoring. A reasonable wavelet domain mapping of the counties in New Mexico was found in Section 3.3.1 and this mapping is shown in Figure 3.5. Appropriate control limits were also determined for each control chart over the monitoring period from October 1977 to December 1991 through simulation. The baseline incidence rate estimates were used to determine the control limits that would lead to an in-control ARL of 200 for each control chart, so

that they could be compared on equal footing. Since the population of females in each county of New Mexico generally increased over time, the control limits required to achieve an ARL of 200 were expected to change as the population size changed. The population estimates provided in the SEER data set were provided for one-year periods. Therefore, simulations were used to search for the appropriate control limit for each control chart over each one-year period, where the population was assumed to remain constant at the estimated value provided in the SEER data set. When the simulations were performed, however, the control limits for the one-year time intervals within each control chart were comparable, regardless of the population differences. As a result, constant control limits were used in the charts over the entire monitoring period, which led to an approximate in-control ARL of 200 for each chart. The control limits used for the chi-square, MEWMA, and Weighted χ^2 control charts were 125, 78.5, and 9.46, respectively. Using constant control limits simplified the analysis because the CUSUM values from previous time intervals did not need to be recomputed to account for the varying population sizes and obtain the CUSUM value for the current year.

Once the baseline rate estimates, wavelet domain mapping, and control limits were determined, the female lung cancer incidences in New Mexico were monitored from October 1977 to December 1991. The chi-square, MEWMA, and Weighted χ^2 control charts for this analysis are shown in Figures 4.6, 4.7, 4.8, respectively. All three control charts signaled quickly, although the MEWMA and Weighted χ^2 control charts signaled first for time interval 13, which corresponds to October 1978. The chi-square control chart signaled soon after the MEWMA and Weighted χ^2 control charts for time interval 20, which corresponds to May 1979.

When the MEWMA and Weighted χ^2 control charts signaled for the month of October in 1978, the SWS and AIC diagnostics were used to determine the reduced model that would adequately describe the changes from baseline in the incidence rate surface. The incidence rate estimates from these models were then used to create the shaded wavelet and geographical maps of New Mexico in Figures 4.9 and 4.10, which can be used to identify subregions with increased lung cancer incidence rates that could be part of a disease cluster. The changes in estimated incidence rates from baseline determined using the full model are shown in the shaded wavelet and geographical maps of Figure 4.11. Each of the diagnostic plots shows that the incidence rate increased in

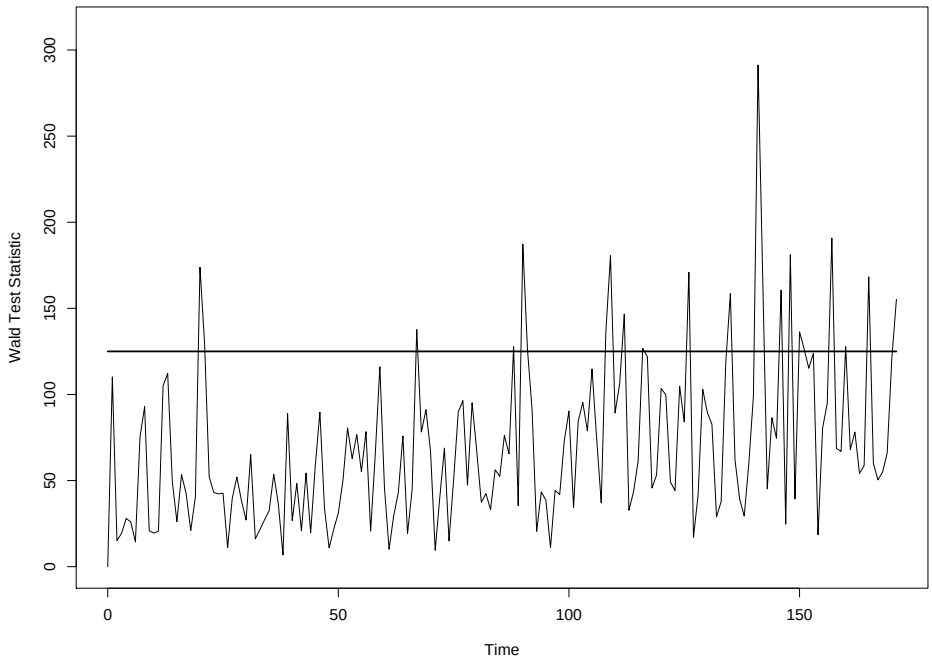


Figure 4.6: Chi-square control chart for the female respiratory lung cancer case study

Lincoln, Chaves, and Otero counties, and that the increased rate was well above the baseline value in each case.

The National Cancer Institute's historical trends tool on the State Cancer Profiles website, statecancerprofiles.cancer.gov, shows that incidences of female respiratory lung cancer in New Mexico increase over the time period when these data were collected. Therefore, based on this analysis, it seems that the wavelet-based surveillance method would have performed well if it had been used for real-time surveillance of these data from 1977 to 1991. The diagnostics were also helpful for identifying the subregions with increased incidence rates and clearly indicated the location of a disease cluster at the time of the first signal. Since the female lung cancer rates have been shown to increase over the entire region during this time period, new areas of increased incidence would be identified by these diagnostics at the time of subsequent signals of the control charts. Taking this into consideration, the cluster related to lung cancer incidences found based

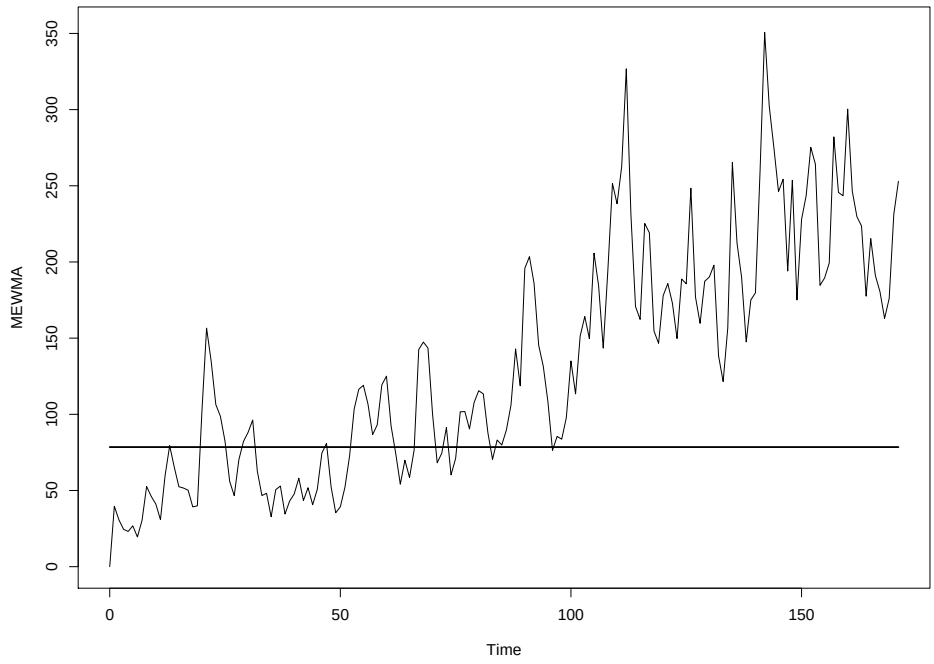


Figure 4.7: MEWMA control chart for the female respiratory lung cancer case study

on the first signal can be thought of as an initial development of increased incidence rates within New Mexico.

Note that in this application, the MEWMA and Weighted χ^2 control charts were not reset after they first signaled in October of 1978, and these charts continued to signal after that time. Typically when control charts are used in manufacturing settings, if a control chart signals, an effort is made to determine the cause of the signal, and to make a process modification so that the monitored parameter returns to its target value before monitoring continues. Once the issue that caused the signal is identified and corrected, control charts that use information from previous observations, such as the MEWMA or CUSUM chart, are reset to their initial values. The control charts used in the wavelet-based method were not reset in this manner because there was no process employed to reduce the incidence rates in the subregions involved in the cluster at the time the signals occurred. Real-time surveillance methods, like this

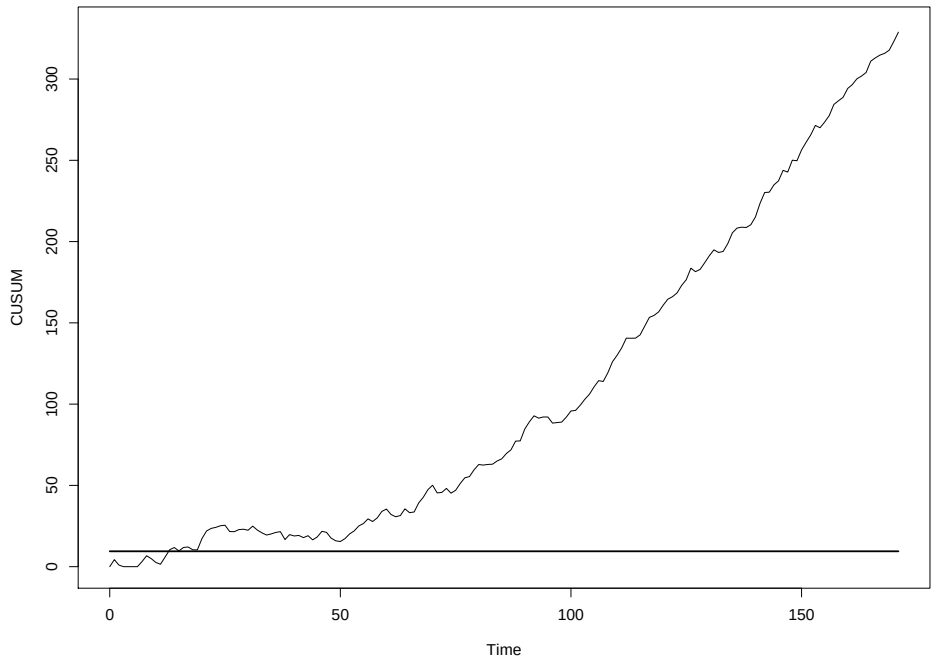
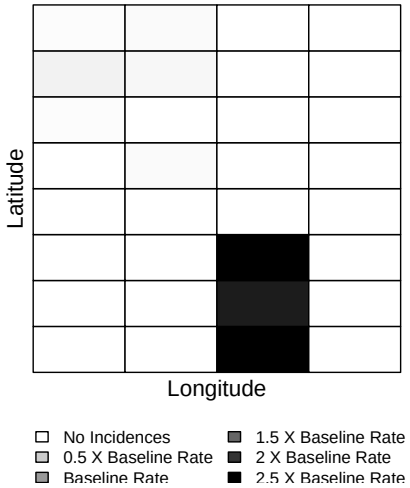


Figure 4.8: Weighted χ^2 control chart for the female respiratory lung cancer case study

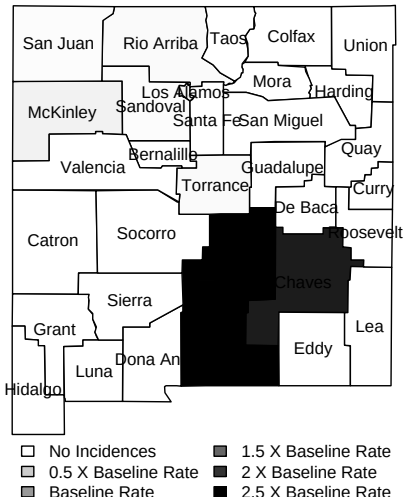
wavelet-based method, are being developed so that monitoring can be done as new data become available and so that preventative action can be taken at the time of a signal. Based on the understanding that the female lung cancer rates increased over the monitoring period and the MEWMA and Weighted χ^2 control charts were not reset, these control charts indicate that the incidence rates continued to increase after the first signal because the values plotted on these charts continue to increase overall for the remainder of the monitoring period.

4.4 Summary and Discussion

Through the evaluation of the wavelet-based disease surveillance method, several factors were shown to have an influence on in-control and out-of-control ARL performance. These factors include the baseline incidence count, the number of subregions within the

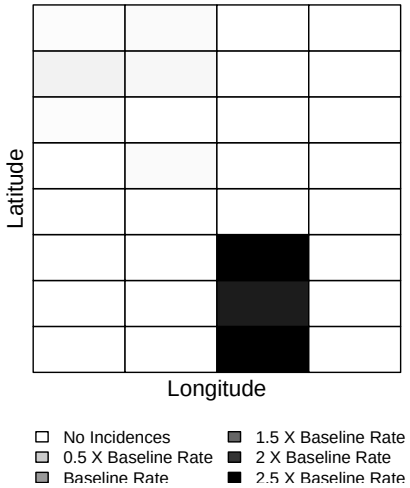


Time 13
(a)

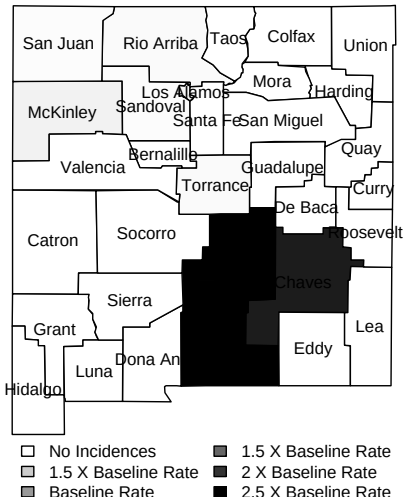


Time 13
(b)

Figure 4.9: Ratios of estimated incidence rates to baseline for the female respiratory lung cancer case study in the (a) wavelet domain; (b) geographical region using the SWS reduced model

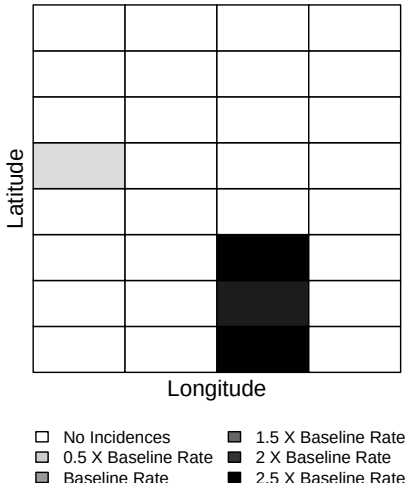


Time 13
(a)

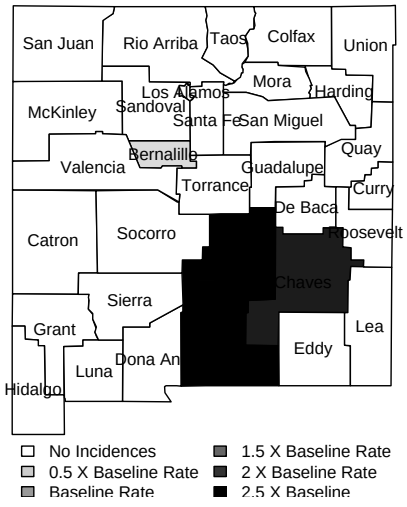


Time 13
(b)

Figure 4.10: Ratios of estimated incidence rates to baseline for the female respiratory lung cancer case study in the (a) wavelet domain; (b) geographical region using the AIC reduced model



Time 13
(a)



Time 13
(b)

Figure 4.11: Ratios of estimated incidence rates to baseline for the female respiratory lung cancer case study in the (a) wavelet domain; (b) geographical region using the full model

region being monitored, the control chart selected, and the Poisson regression model used. In the in-control ARL evaluation of the wavelet-based method, the model used and the baseline incidence counts had the largest impact on performance. When the control limits of each control chart were set so that the nominal ARL was equal to 200, based on the approximate distribution of the monitored parameters, the control charts generally had true in-control ARLs closer to 200 when the identity link model was used as opposed to the canonical link model. With both models, the true in-control ARLs for all control charts were furthest from 200 when the baseline incidence count was low, but they approached 200 as the baseline count increased. The MEWMA and Weighted χ^2 control charts had in-control ARLs closest to 200 for low baseline counts when the identity link model was used.

In the evaluation of the out-of-control ARL performance of the control charts in the wavelet-based method, the MEWMA control chart proved to have the best performance for more combinations of subregion size, baseline count, and model across all out-of-control scenarios when compared to the other control charts evaluated. This chart performed at its best when it was used with the identity link Poisson regression model. The one-sided and two-sided subregion control charts also performed well, specifically for the out-of-control scenarios where the cluster sizes were small. In several cases when small clusters were present, these control charts outperformed the MEWMA control chart when using the canonical link model. Overall, the performance for the chi-square and Weighted χ^2 control charts was poor when compared to the other charts, although there were out-of-control scenarios where these charts performed best. In scenarios where there were large clusters present, the baseline count was equal to 1, and the canonical link model was used, the chi-square control chart performed best. In other cases where there were large clusters present, the Weighted χ^2 control chart had the best out-of-control ARL performance. In general, even though both the chi-square control chart and the Weighted χ^2 control chart performed poorly, the Weighted χ^2 chart outperformed the chi-square chart.

The wavelet-based method performed well as a whole when applied in a real-world scenario. In the case study where female lung cancer incidences in New Mexico were monitored, the incidence rates were known to increase over time. The chi-square, MEWMA, and Weighted χ^2 control charts using the identity link model all signaled

over the monitoring period as expected, indicating that the incidence rate surface had changed. Still, the MEWMA and Weighted χ^2 control charts signaled more quickly than the chi-square chart. The SWS and AIC diagnostic tools, used to determine the location of a cluster after a signal, also worked well for identifying the area in New Mexico where the incidence rates had increased after the MEWMA and Weighted χ^2 control charts produced their first signals.

Although the Weighted χ^2 control chart was able to detect an increase in the incidence rates of several subregions quickly in the case study, it did not perform as well as expected based on out-of-control performance. This control chart was expected to outperform both the chi-square and MEWMA control charts based on the use of a weighted χ^2 statistic, which used a weighting scheme that put more emphasis on the low resolution wavelet coefficients. In most cases, however, the MEWMA control chart had better out-of-control ARL performance when compared to the Weighted χ^2 control chart. This may be due to the fact that the weighted χ^2 statistic is based on a Wald statistic, which may not lead to good performance in this application. It may also be due to the weighting scheme applied to the statistic. An alternative weighted statistic developed by Spitzner and Marshall (2008) could be used in place of the weighted χ^2 statistic monitored in the Weighted χ^2 control chart to see if the performance of the chart can be improved.

Chapter 5

Summary and Discussion of Future Research

The primary goal of the research presented in the previous chapters was to show the development, evaluation, and improvement of spatio-temporal disease surveillance methods for the detection of chronic disease clusters. The research toward this goal consists of three main contributions. In Chapter 2, the method of Rogerson (2001) was evaluated based on in-control average run length (ARL) performance. This method was designed to detect disease clusters in a geographical region when data with no aggregation in space and time are available. Rogerson (2001) proposed the use of a cumulative sum (CUSUM) chart for monitoring a local Knox statistic to identify forming clusters of disease by detecting space-time interaction. Identifying clusters of disease by monitoring space-time interaction is advantageous because this strategy does not require the use of a known baseline incidence rate. When evaluating this method, the in-control ARL performance of the CUSUM chart was shown to be influenced by many factors, including the population density within the monitored region, changes in population density over time, the region shape, and the space and time thresholds selected. Because the there are so many factors that can impact the in-control ARL of this chart, the control limit must be determined through simulation in each new application to achieve a specified in-control ARL. In addition, the standard normal distribution assumed for the local Knox statistic did not approximate the true distribution of this statistic well. Contending with the poor approximation also requires that the control limit be found by simulation for any new application. Ultimately, this method could

not be recommended, regardless of the issues with in-control ARL performance and approximate distribution, because the method does not incorporate changes in the population over time. This can lead to the false detection of a disease cluster since an increase in population over time mimics the formation of a disease cluster.

In Chapter 3, a new spatio-temporal disease surveillance method was developed based on the use of Haar wavelet functions. This wavelet-based method assumes that the data available are aggregated in both space and time. In this method, the subregions within a geographical region are mapped to the Haar wavelet domain. Then a Poisson regression model is used to model the incidence rate surface over the wavelet domain by using the Haar wavelet functions as regressors in the model. The coefficients of this model are monitored through the use of control charts to detect changes in the incidence rate surface from baseline over time. Once a signal occurs indicating a change in the incidence rate surface, diagnostics are used to determine the location and shape of the disease cluster or clusters present.

The wavelet-based method has features that make it unique when compared to other methods developed for prospective spatio-temporal disease monitoring. The wavelet-based method is different than other methods for prospective disease surveillance because it takes a profile monitoring approach to disease cluster detection. This method monitors an incidence surface or profile over time to determine if clusters are present by using a control chart. The wavelet-based method is also unique because of the use of wavelets as regressors in the Poisson regression model. The use of wavelets is important because of their multiresolution. The multiresolution of wavelets allows for the incidence surface to be partitioned into both small and large areas, which makes it possible to detect disease clusters of any size and shape within a geographical region. The multiresolution of wavelets also made it possible to develop a statistic that can place more emphasis on the detection of disease clusters of a specified size. Each wavelet coefficient is associated with a particular resolution that partitions the region into areas of a specific size. The weighted χ^2 statistic was developed so that the contribution of the wavelet coefficients of a specified resolution can be weighted more heavily. When this statistic is used in a control chart, the method has more power to detect clusters at the specified resolution. No other disease surveillance method is able to focus statistical power on the detection of disease clusters of a certain size.

The wavelet-based method is also able to handle the issues of multiple testing and changes in the population and covariates within a geographical region, where some other methods fall short. This method avoids multiple testing by monitoring the parameters of the disease incidence rate surface using a multivariate control chart. This allows for the detection of a change in the disease incidence rate in any subregion by doing one test for each time interval. Diagnostics are used to determine the locations of incidence rate changes only when the control chart signals. The wavelet-based method also allows for the population and covariates to change within each subregion over time without influencing the performance of the method. This is done by including an offset for population and an offset for covariates in the Poisson regression model using the canonical link function. When the identity link function is used, this is accomplished by adjusting the baseline counts for the changes in population and covariates for each time interval.

While the wavelet-based method has many attributes that make it useful for the detection of disease clusters, this method also has some shortcomings. One is that the baseline rates must be known or estimated for each subregion within the geographical region being monitored. In some cases, this information will not be available, making the method difficult to apply. Another issue is related to the mapping of the geographical region to the wavelet domain. In many cases, a reasonable mapping can be found, but when the mapping algorithm was evaluated there were a few cases where the algorithm failed to find a reasonable mapping. This could be due to the setting of the parameter κ used in the search algorithm. It could also be due to the search algorithm selected, which was a simple local search algorithm. Further investigation is needed to determine the cause. The SWS and AIC diagnostics, used to locate the disease cluster or clusters following a signal, can also pose a problem. When a signal occurs due to an increase in the incidence rates that has been accumulating over time, it can sometimes be difficult to determine the location of a cluster when these diagnostics are used because they only use information from the current observation. These diagnostics could be improved if they were modified to incorporate information from prior observations like the control charts. One other aspect of the wavelet-based method that may be undesirable in some applications, is that it detects both increases and decreases in the incidence rates. Some practitioners may prefer a method that only detects increases in

the incidence rates over time. Therefore, it is important to explore potential ways to modify the wavelet-based method so that it can be used for detecting only increases in the incidence rates.

In Chapter 4, the wavelet-based disease surveillance method was evaluated based on in-control and out-of-control ARL performance. The evaluation of in-control ARL performance showed that when the baseline counts are low, the true in-control ARLs of the control charts used in this method are far from the nominal value determined by using the approximate distribution assumed for the coefficients. As the baseline rates increase, the in-control ARLs for the control charts approach the nominal value. Therefore, when the baseline counts are low, simulation is needed to determine control limits that achieve a specified in-control ARL. When the baseline counts increase, however, the approximate distribution of the coefficients can be used to determine the control limits. In the evaluation of out-of-control ARL performance, the multivariate exponentially weighted moving average (MEWMA) control chart using the identity link Poisson regression model proved to have the best ARL performance in a majority of the out-of-control scenarios considered. The Weighted χ^2 control chart did not perform as well as expected in many scenarios, although it did have low out-of-control ARLs when the cluster size covered at least half of the region under surveillance. In addition to being able to detect large clusters quickly, the Weighted χ^2 control chart was also expected to have good performance for the detection of small clusters based on the weighting scheme used in the weighted χ^2 statistic. Although the performance of this chart was not as good as expected, this control chart can be modified to potentially improve its performance. The weighting scheme of the weighted χ^2 statistic can be adjusted or a new weighted statistic could be used with this control chart to see if the out-of-control ARL performance is improved when small clusters are present.

The development of disease surveillance methods continues to be an important area of research. There are many monitoring scenarios that this research does not cover and where more work is needed. Specifically, there is a need for the development of methods that can be used for monitoring contagious diseases. These methods are typically much more complex because they require the modeling of spatial and temporal autocorrelation.

Appendix A

Variance of the Local Knox Statistic

Suppose that for a fixed spatial location at observation i, $\{N_{st}(i)_j : j = 1, 2, \ldots, n\}$ are hypergeometrically distributed random variables with parameters $n - 1$, $n_s(i)$, and $n_t(j)$. Further suppose that the random variable $N_{st}(i)$ is equal to $N_{st}(i)_j$ with probability $1/n$ for $j = 1, 2, \ldots, n$. This can be written as,

$$N_{st}(i) \quad = \quad \sum_{j=1}^{n} A_j N_{st}(i)_j, \tag{A.1}$$

where

$$P(A_j = 1) \quad = \quad \begin{cases} \frac{1}{n} & \text{for } j = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} . \tag{A.2}$$

Then it follows that

$$P\left(N_{st}(i) = n_{st}(i)\right) = \sum_{j=1}^{n} \frac{1}{n} P_j\left(n_{st}(i)\right), \tag{A.3}$$

where P_j is the probability mass function for the hypergeometric random variable with parameters $n - 1$, $n_s(i)$, and $n_t(j)$.

The variance of $N_{st}(i)$ is

$$\begin{aligned} V\left[N_{st}(i)\right] &= E\left[N_{st}(i)^2\right] - E\left[N_{st}(i)\right]^2 \\ &= \sum_{All\ n_{st}(i)} n_{st}(i)^2 P\left(N_{st}(i) = n_{st}(i)\right) - E\left[N_{st}(i)\right]^2 \\ &= \sum_{All\ n_{st}(i)} n_{st}(i)^2 \sum_{j=1}^{n} \frac{1}{n} P_j\left(n_{st}(i)\right) - E\left[N_{st}(i)\right]^2 \\ &= \sum_{j=1}^{n} \frac{1}{n} \sum_{j=1}^{n} n_{st}(i)^2 P_j\left(n_{st}(i)\right) - E\left[N_{st}(i)\right]^2 \\ &= \frac{1}{n} \sum_{j=1}^{n} E\left[N_{st}(i)_j^2\right] - E\left[N_{st}(i)\right]^2 \\ &= \frac{1}{n} \sum_{j=1}^{n} \left[V\left[N_{st}(i)_j\right] + E\left[N_{st}(i)_j\right]^2\right] - E\left[N_{st}(i)\right]^2. \end{aligned} \tag{A.4}$$

Since $N_{st}(i)_j$ has a hypergeometric distribution with parameters $n-1$, $n_s(i)$, and $n_t(j)$, then the expectation and variance of $N_{st}(i)_j$ are

$$E\left[N_{st}(i)_j\right] = \frac{n_s(i) n_t(j)}{n-1} \tag{A.5}$$

and

$$V\left[N_{st}(i)_j\right] = \left[\frac{n_s(i) n_t(j)}{n-1}\right] \left[\frac{\left(n-1-n_t(j)\right)\left(n-1-n_s(i)\right)}{(n-1)(n-2)}\right], \tag{A.6}$$

respectively.

By substituting the expressions for the mean and variance of $N_{st}(i)_j$ in equations (A.5) and (A.6) into equation (A.4) and using the fact that $\sum_{j=1}^{n} n_t(j) = 2n_t$, the expression for the variance of $N_{st}(i)$ becomes

$$\begin{aligned} V\left[N_{st}(i)\right] &= \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\left(n-1-n_s(i)\right)}{n(n-1)^2(n-2)} \\ &\quad + \frac{n_s(i)^2 \sum_{i=1}^{j} n_t(j)^2}{n(n-1)} - E\left[N_{st}(i)\right]^2. \end{aligned} \tag{A.7}$$

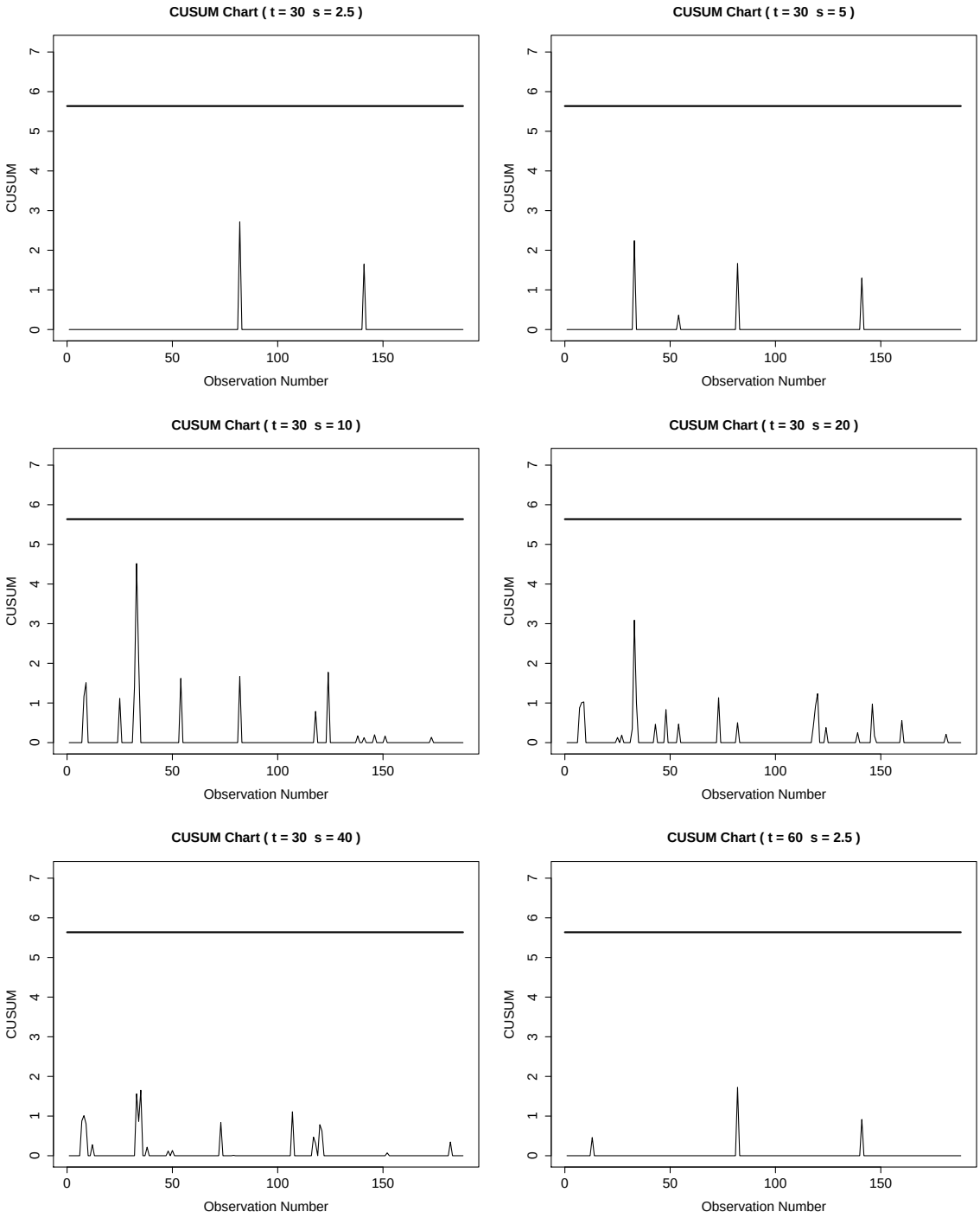
Finally, by replacing $E\left[N_{st}(i)\right]$ in equation (A.7) with the expression for $E\left[N_{st}(i)\right]$ given in equation (2.2), the variance of $N_{st}(i)$ is

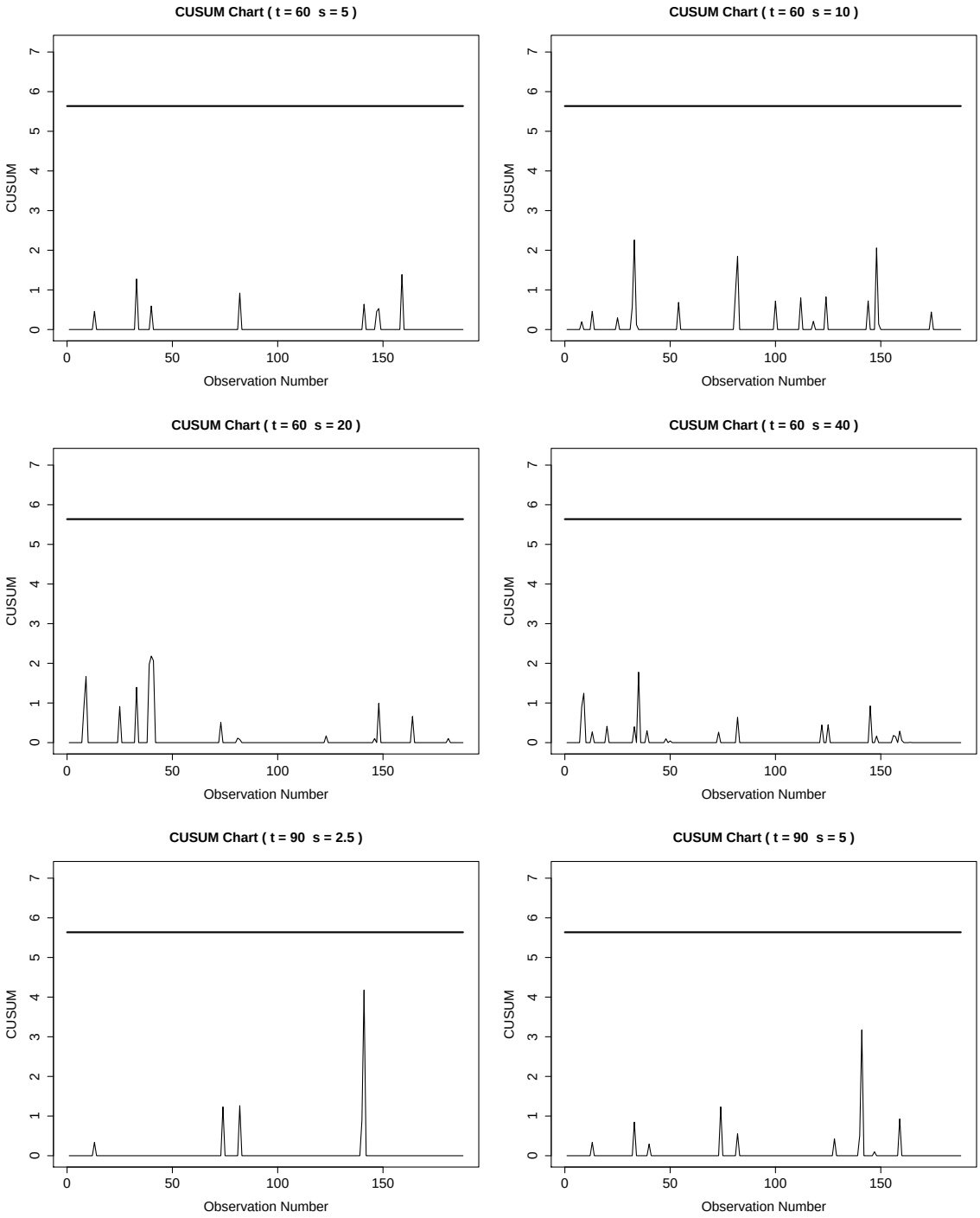
$$\begin{aligned} V[N_{st}(i)] &= \frac{\left[2(n-1)n_t - \sum_{j=1}^{n} n_t(j)^2\right] n_s(i)\left(n-1-n_s(i)\right)}{n(n-1)^2(n-2)} \\ &\quad + \frac{n_s(i)^2 \sum_{j=1}^{n} \left[n_t(j) - \frac{2n_t}{n}\right]^2}{n(n-1)^2}. \end{aligned} \tag{A.8}$$

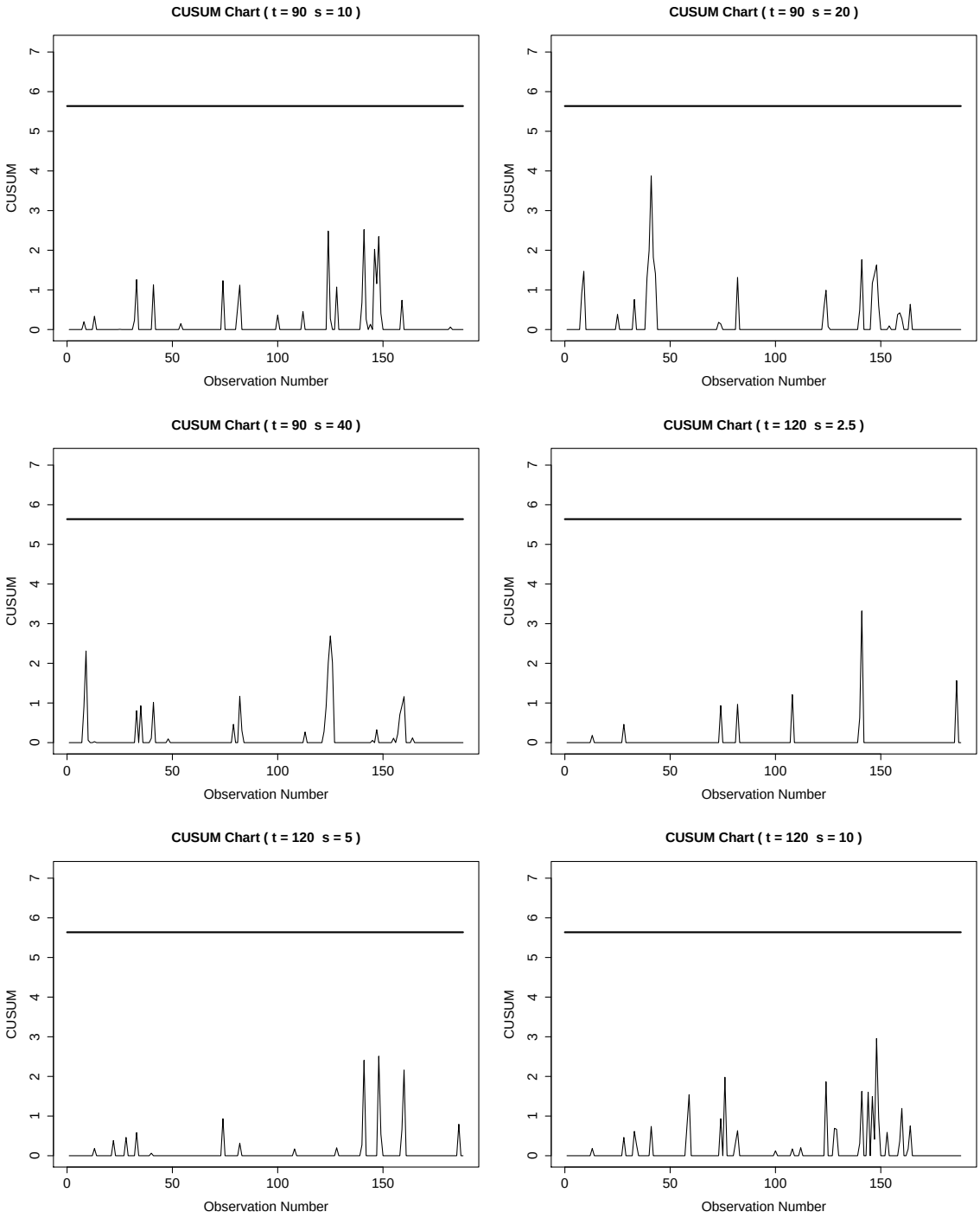
Appendix B

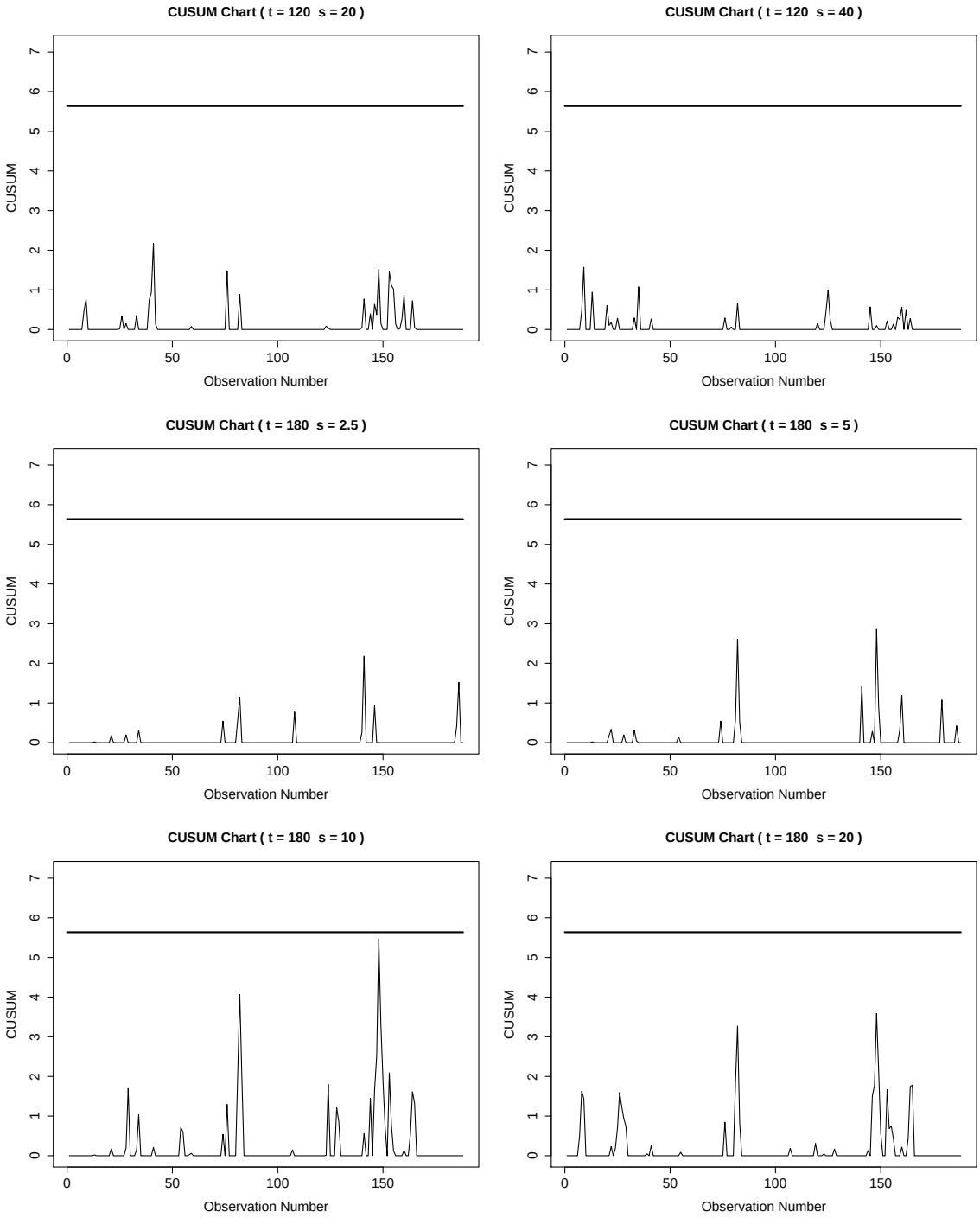
CUSUM Charts for the Local Knox Monitoring Method Application

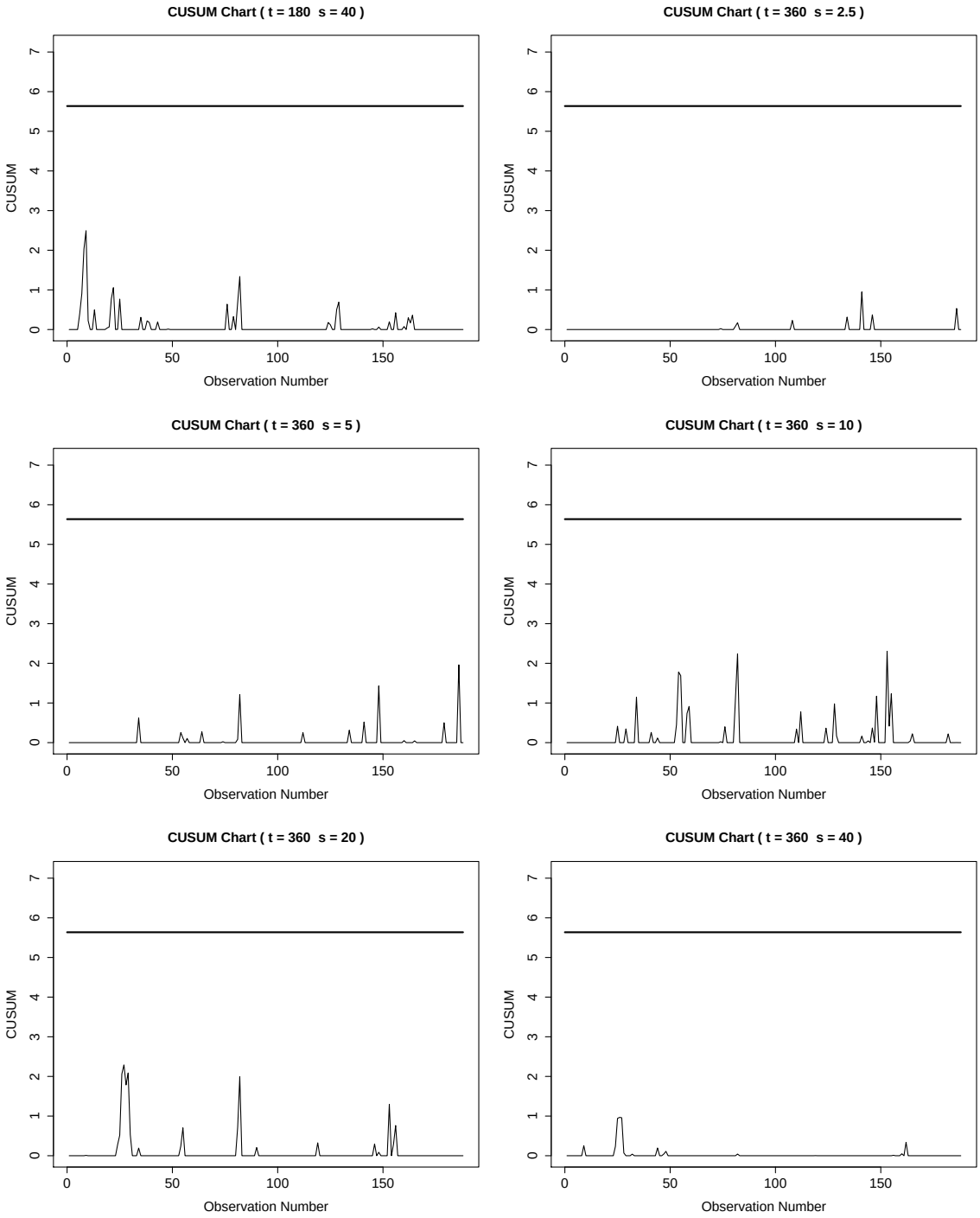
Rogerson (2001) developed the local Knox monitoring method and used the method to detect space-time clusters of Burkitt's lymphoma in the West Nile District of Uganda. The original application of this method, presented in Rogerson (2001), was incorrect due to computational errors and errors in the data set used. The approximate variance of the local Knox statistic in equation (2.4) was also used in the analysis instead of the exact variance in equation (2.3), which affected the results. This appendix shows the corrected control charts for the Burkitt's lymphoma application, using the exact variance of the local Knox statistic in equation (2.3). A total of 30 space and time threshold combinations where used in the analysis, and the CUSUM control charts for all of these combinations are shown.


CUSUM Chart (t = 30 s = 2.5)
CUSUM Chart (t = 30 s = 5)
CUSUM Chart (t = 30 s = 10)
CUSUM Chart (t = 30 s = 20)
CUSUM Chart (t = 30 s = 40)
CUSUM Chart (t = 60 s = 2.5)
CUSUM
Observation Number
0
1
2
3
4
5
6
7
0
50
100
150


CUSUM Chart (t = 60 s = 5)
CUSUM Chart (t = 60 s = 10)
CUSUM Chart (t = 60 s = 20)
CUSUM Chart (t = 60 s = 40)
CUSUM Chart (t = 90 s = 2.5)
CUSUM Chart (t = 90 s = 5)
CUSUM
Observation Number
0
1
2
3
4
5
6
7
0
50
100
150


CUSUM Chart (t = 90 s = 10)
CUSUM Chart (t = 90 s = 20)
CUSUM Chart (t = 90 s = 40)
CUSUM Chart (t = 120 s = 2.5)
CUSUM Chart (t = 120 s = 5)
CUSUM Chart (t = 120 s = 10)
CUSUM
Observation Number
0
1
2
3
4
5
6
7
0
50
100
150

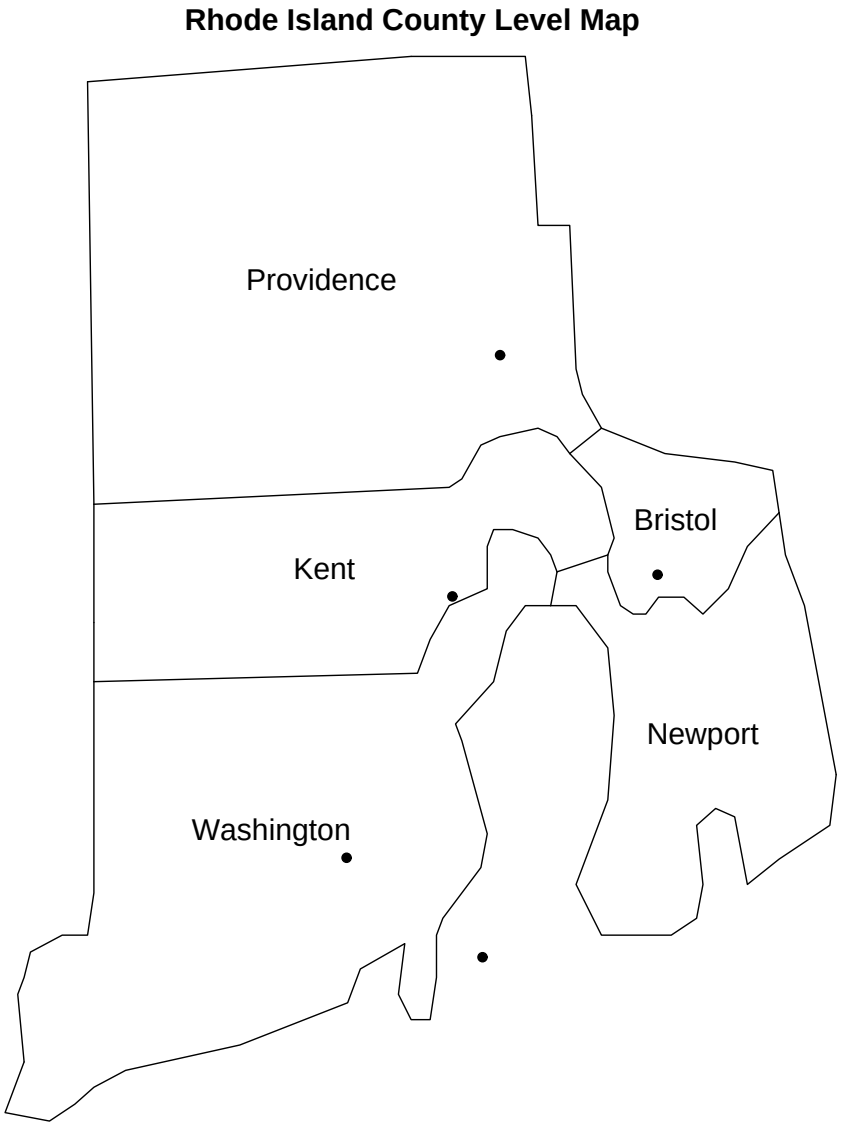

CUSUM Chart (t = 120 s = 20)
CUSUM Chart (t = 120 s = 40)
CUSUM Chart (t = 180 s = 2.5)
CUSUM Chart (t = 180 s = 5)
CUSUM Chart (t = 180 s = 10)
CUSUM Chart (t = 180 s = 20)
CUSUM
Observation Number
0
1
2
3
4
5
6
7
0
50
100
150

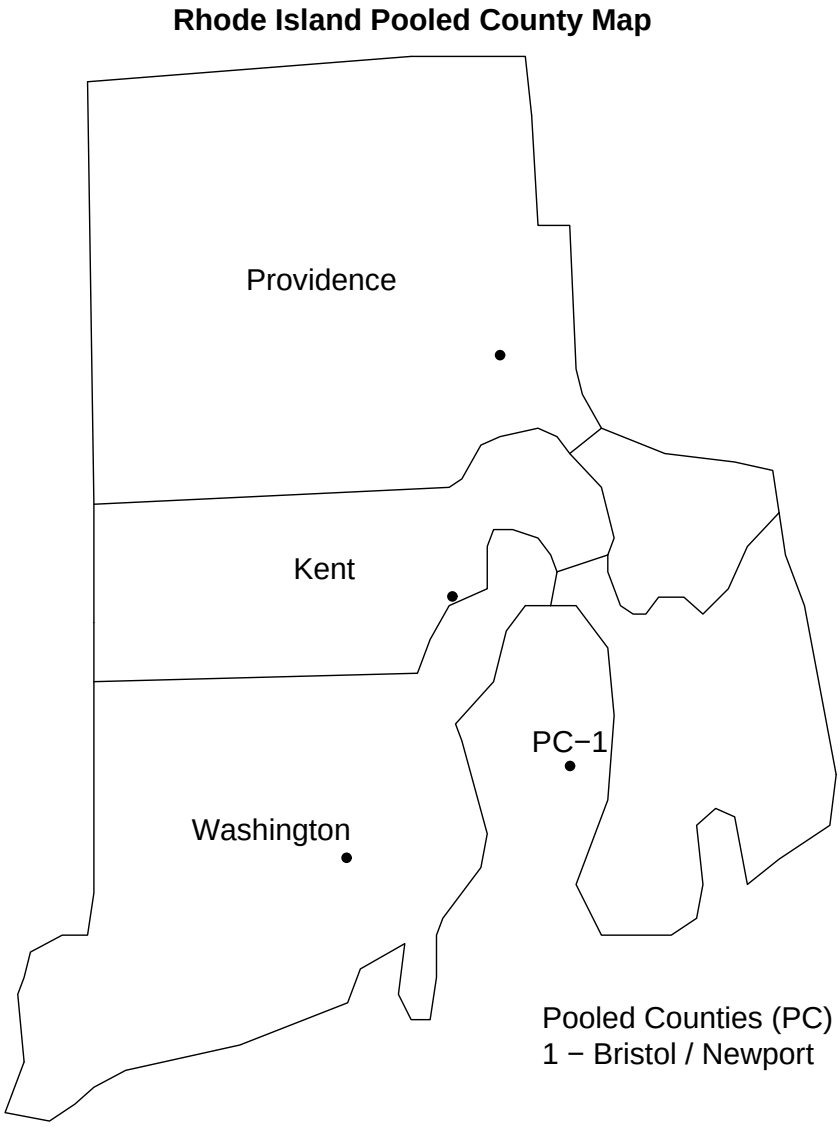

CUSUM Chart (t = 180 s = 40)
CUSUM Chart (t = 360 s = 2.5)
CUSUM Chart (t = 360 s = 5)
CUSUM Chart (t = 360 s = 10)
CUSUM Chart (t = 360 s = 20)
CUSUM Chart (t = 360 s = 40)
CUSUM
Observation Number
0
1
2
3
4
5
6
7
0
50
100
150

Appendix C

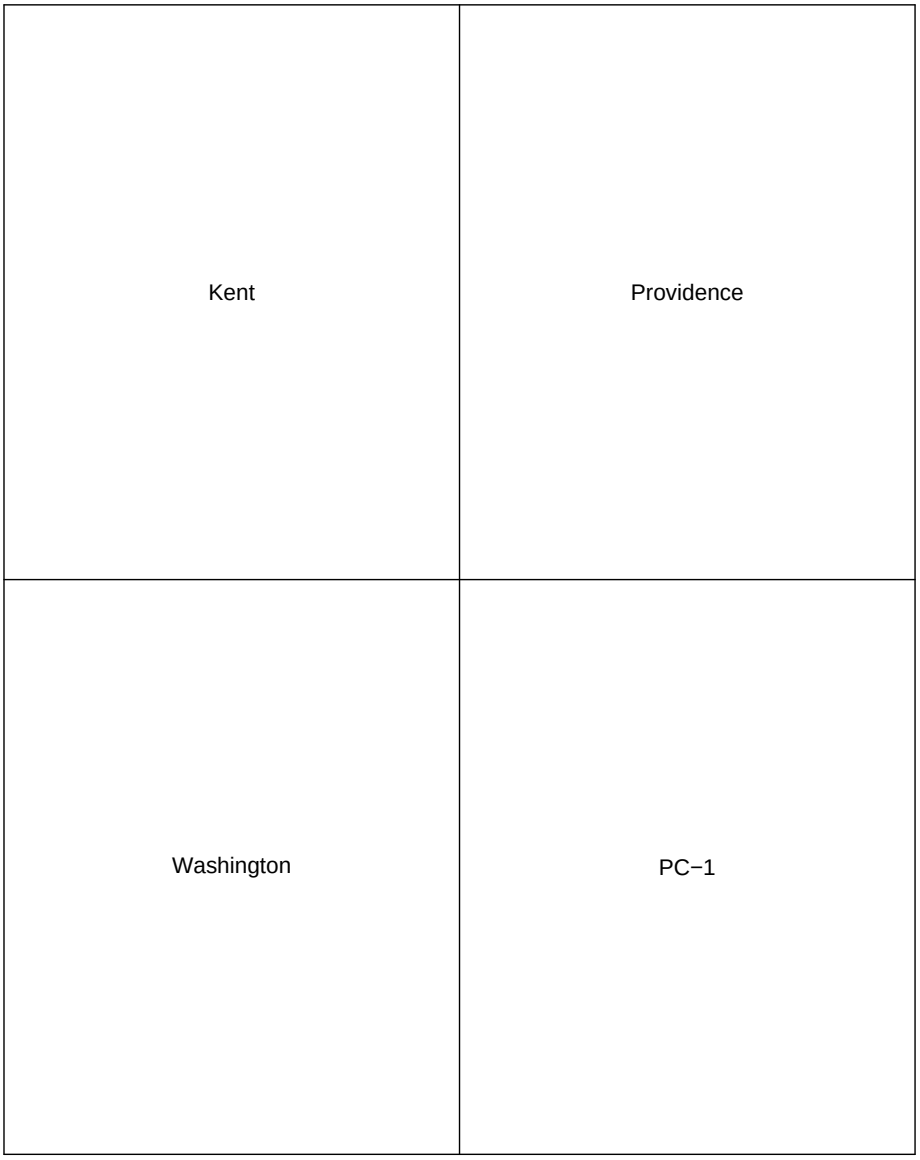
Wavelet Domain Mapping Examples for Ten States

This appendix provides the geographical maps and the wavelet domain mappings found using the simple local search algorithm discussed in Chapter 3, Section 3.3.1 for the ten US states used to evaluate the algorithm. The geographical map, the pooled geographical map for states with pooled counties, and the wavelet domain maps for the distance, direction, and adjacent objective functions are shown for Rhode Island, Connecticut, New Hampshire, Arizona, Maine, Maryland, New Mexico, California, Louisiana, and Ohio, respectively. The county seats used in the calculation of the distance and direction objective functions are indicated on both the geographical and pooled geographical maps by black dots.


Rhode Island County Level Map
Providence
Bristol
Kent
Newport
Washington


Rhode Island Pooled County Map
Providence
Kent
PC−1
Washington
Pooled Counties (PC)
1 − Bristol / Newport

Rhode Island Wavelet Domain Mapping − Distance

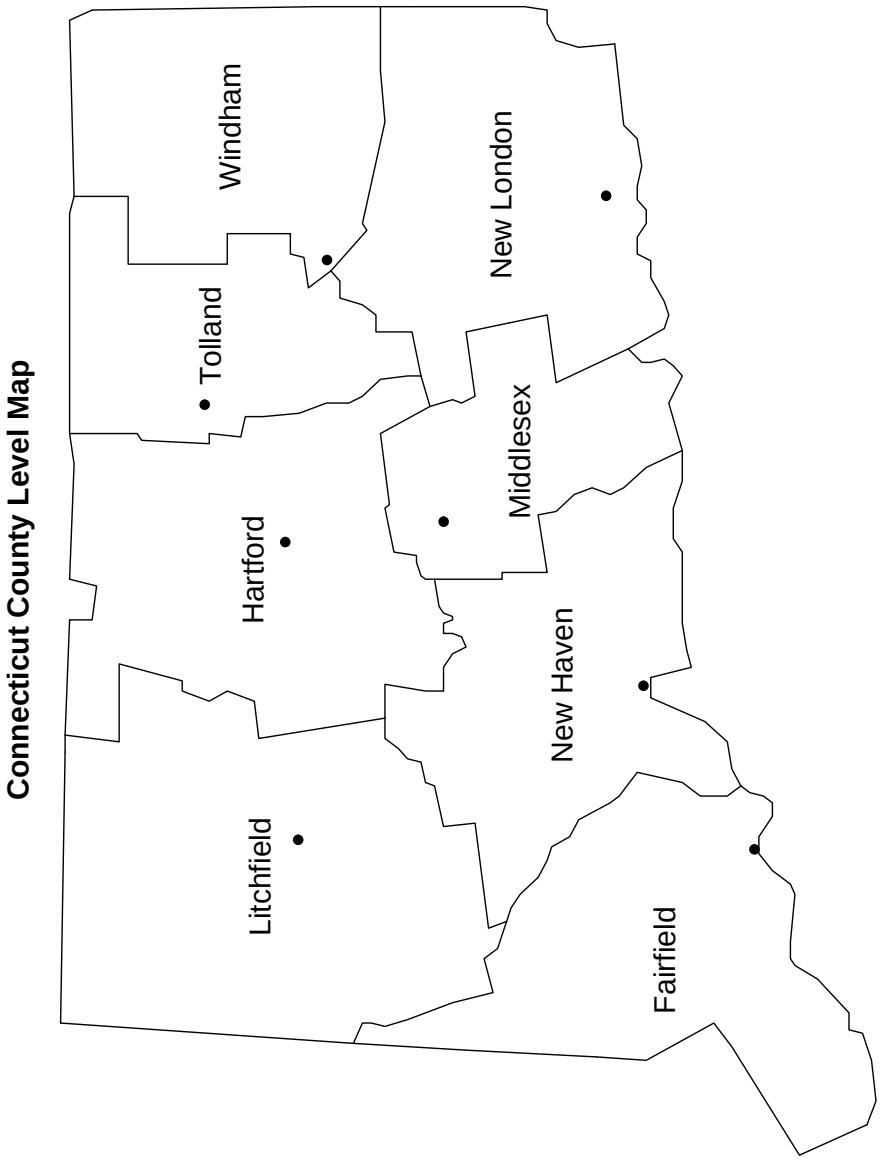


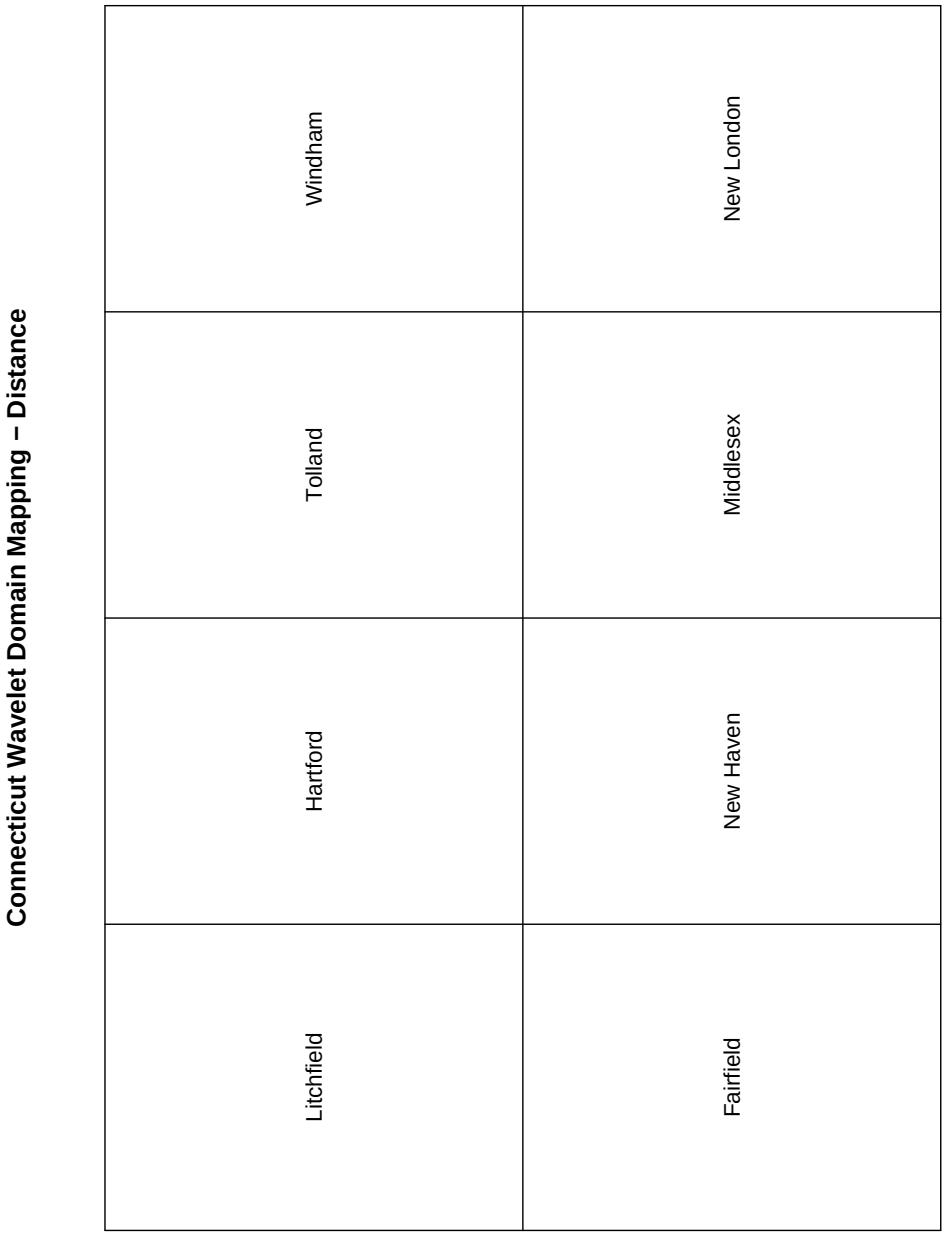
Rhode Island Wavelet Domain Mapping − Direction

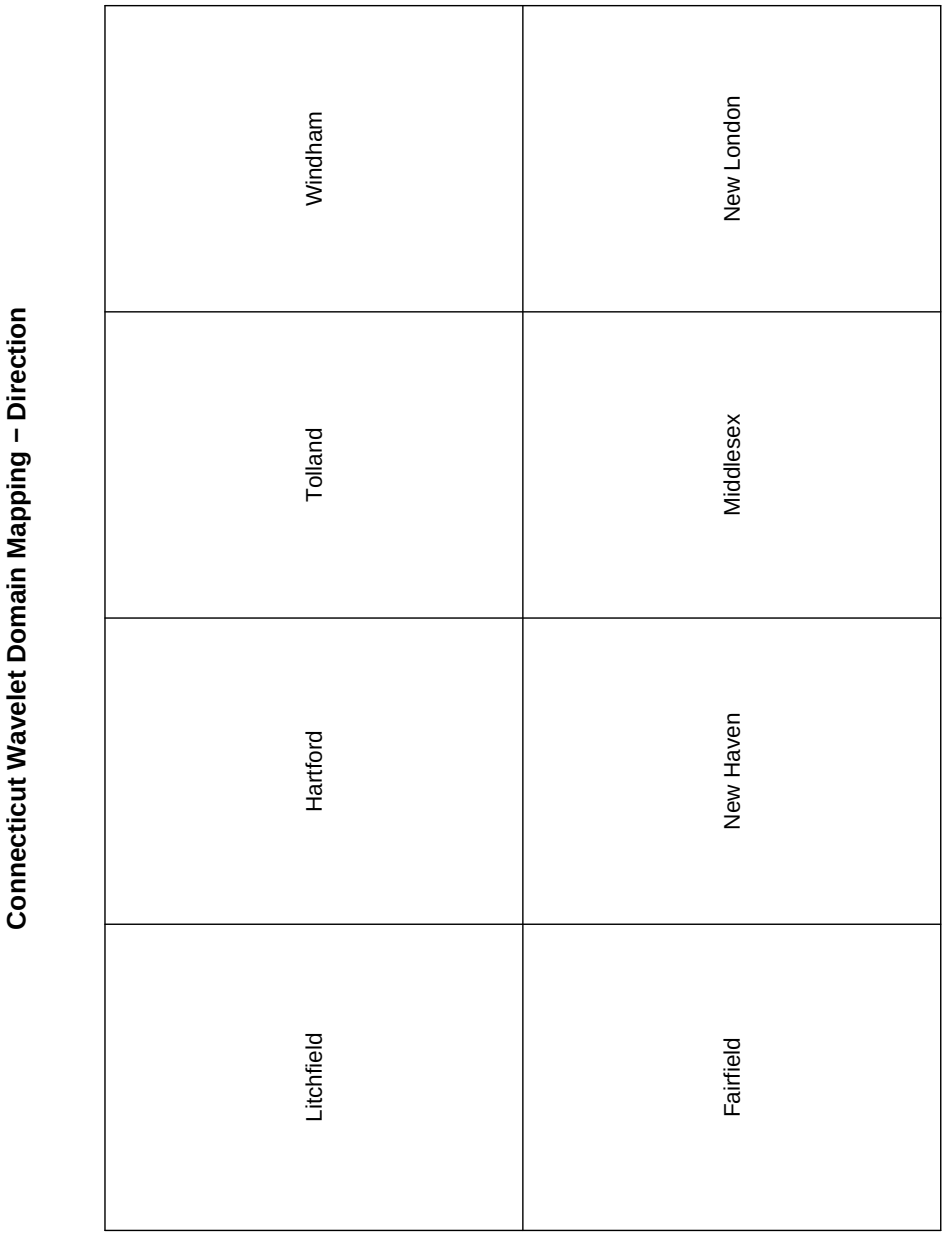
Kent	Providence
Washington	PC−1

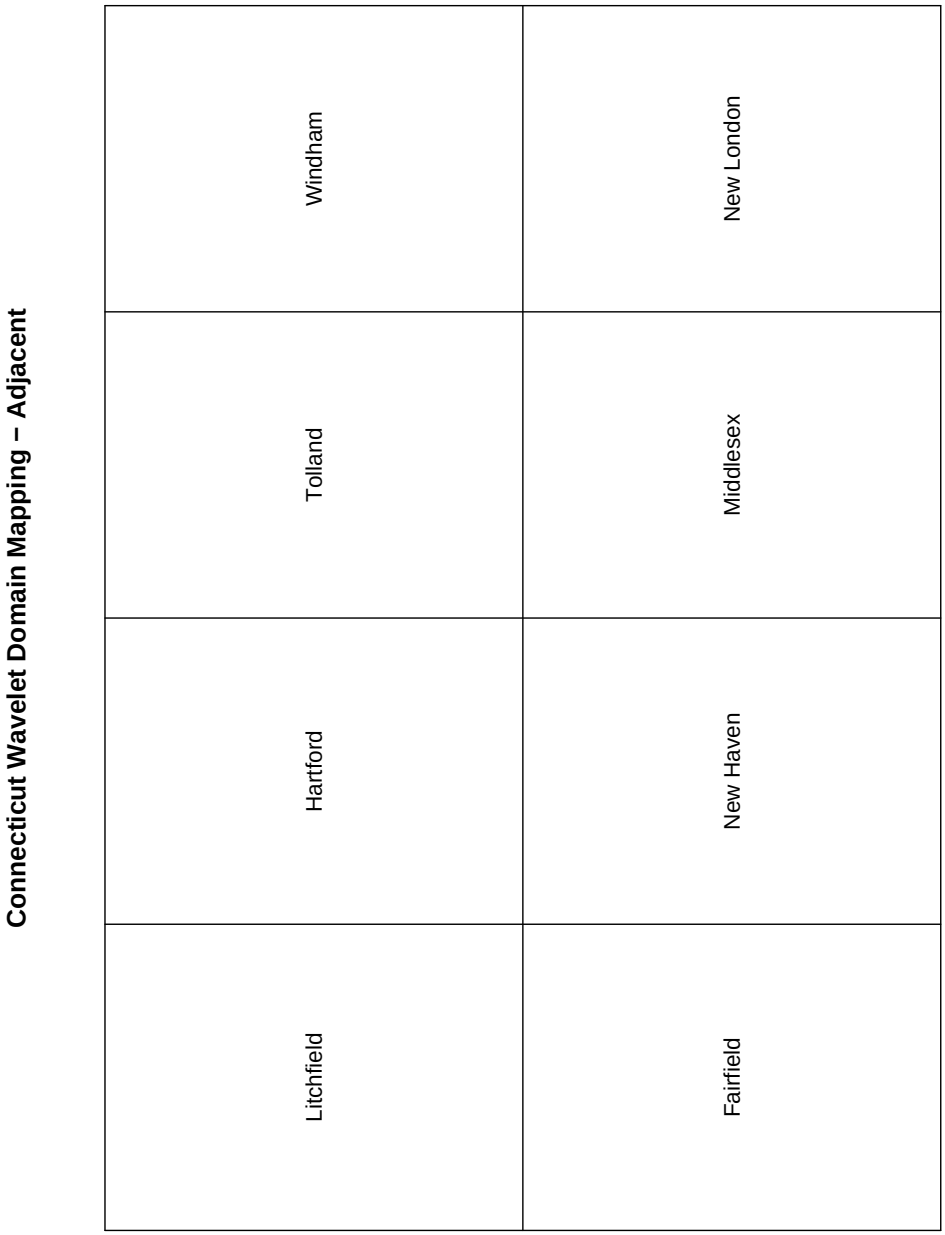
Rhode Island Wavelet Domain Mapping − Adjacent

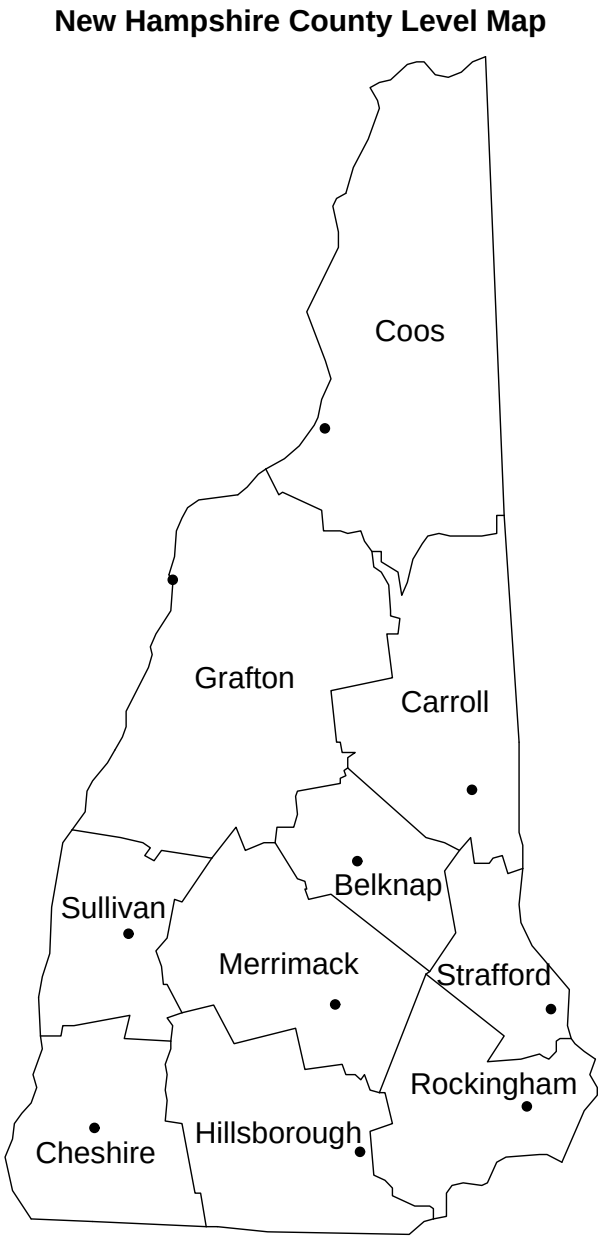
Kent	Providence
Washington	PC−1

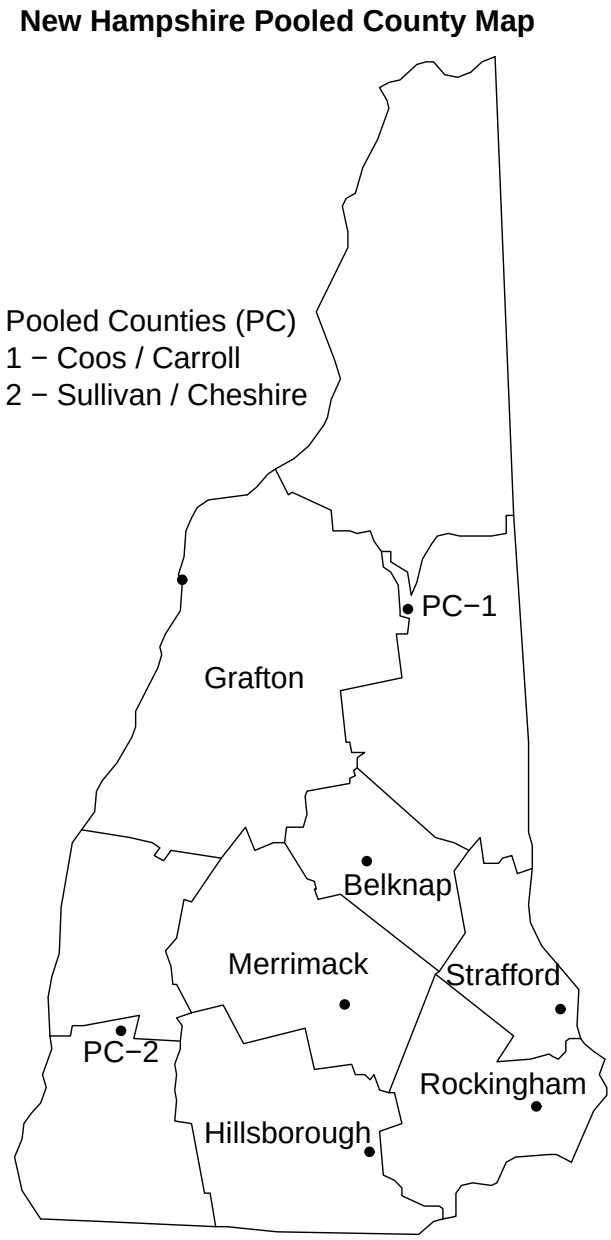

Connecticut County Level Map
Windham
New London
Tolland
Middlesex
Hartford
New Haven
Litchfield
Fairfield


Connecticut Wavelet Domain Mapping – Distance
Windham
New London
Tolland
Middlesex
Hartford
New Haven
Litchfield
Fairfield


Connecticut Wavelet Domain Mapping – Direction
Windham
New London
Tolland
Middlesex
Hartford
New Haven
Litchfield
Fairfield


Connecticut Wavelet Domain Mapping – Adjacent
Windham
New London
Tolland
Middlesex
Hartford
New Haven
Litchfield
Fairfield


New Hampshire County Level Map
Coos
Grafton
Carroll
Belknap
Sullivan
Merrimack
Strafford
Rockingham
Hillsborough
Cheshire


New Hampshire Pooled County Map
Pooled Counties (PC)
1 − Coos / Carroll
2 − Sullivan / Cheshire
PC−1
Grafton
Belknap
Merrimack
Strafford
PC−2
Rockingham
Hillsborough

New Hampshire Wavelet Domain Mapping − Distance

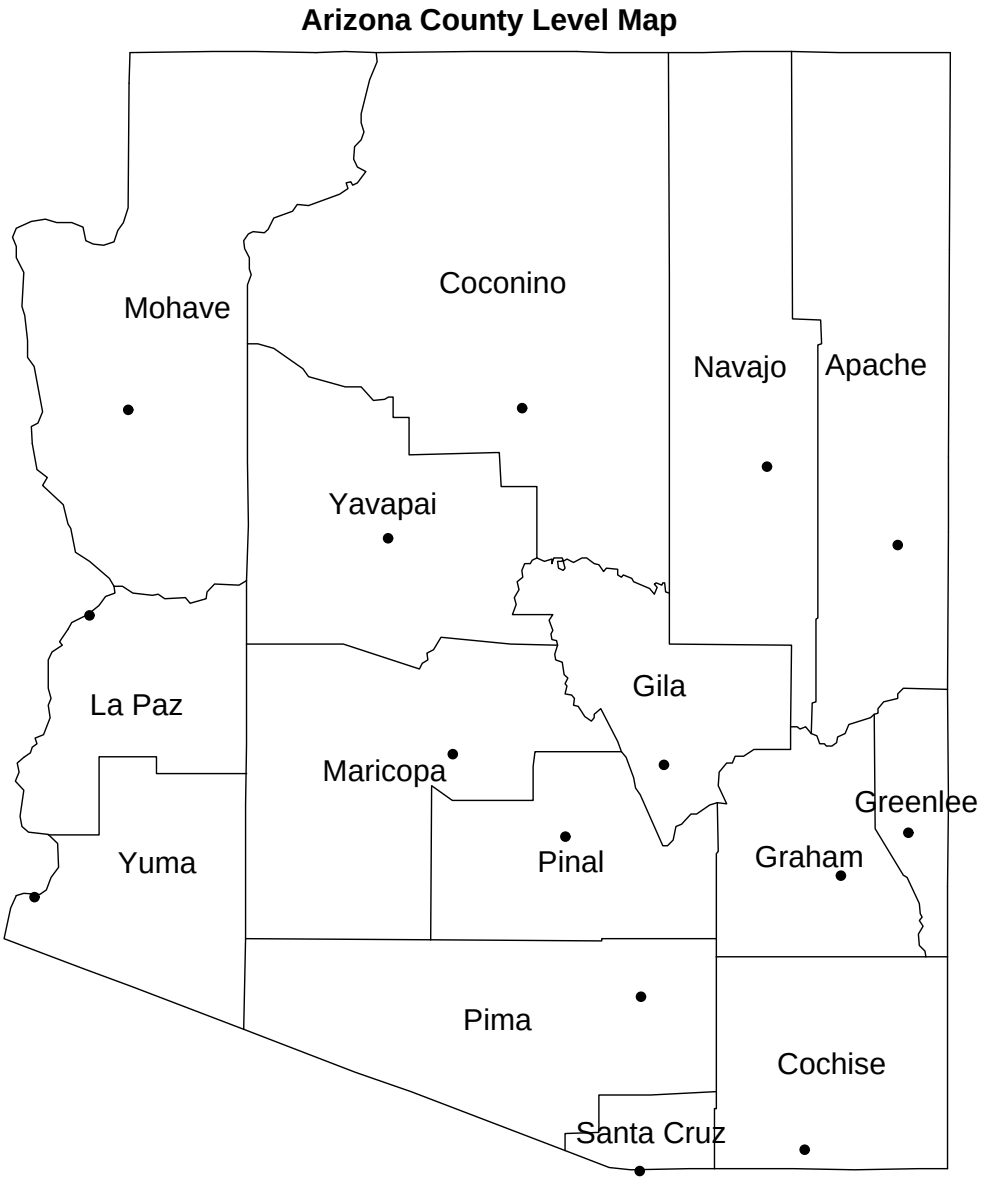
Grafton	Merrimack
PC−1	Strafford
PC−2	Hillsborough
Belknap	Rockingham

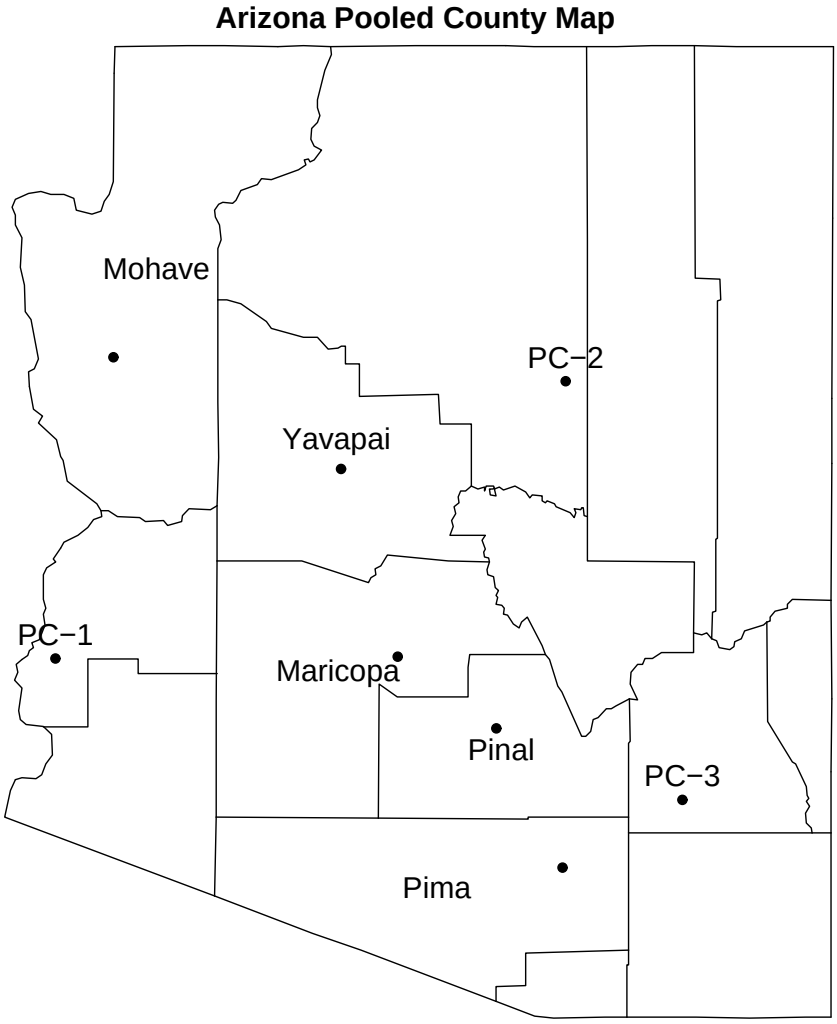
New Hampshire Wavelet Domain Mapping − Direction

Grafton	Belknap
PC−2	Merrimack
PC−1	Strafford
Hillsborough	Rockingham

New Hampshire Wavelet Domain Mapping − Adjacent

Grafton	PC−1
Belknap	Strafford
Merrimack	Rockingham
PC−2	Hillsborough


Arizona County Level Map
Mohave
Coconino
Navajo
Apache
Yavapai
La Paz
Gila
Maricopa
Greenlee
Yuma
Pinal
Graham
Pima
Cochise
Santa Cruz



Pooled Counties (PC)
1 − La Paz / Yuma
2 − Navajo / Coconino
3 − Apache / Greenlee / Graham / Gila / Santa Cruz / Cochise

Arizona Wavelet Domain Mapping − Distance

Mohave	Yavapai	PC−2	PC−3
PC−1	Maricopa	Pinal	Pima

Arizona Wavelet Domain Mapping − Direction

Mohave	Yavapai	PC−2	Pinal
PC−1	Maricopa	Pima	PC−3

Arizona Wavelet Domain Mapping − Adjacent

Mohave	PC−2	PC−3	Pinal
PC−1	Yavapai	Maricopa	Pima


Maine County Level Map
Aroostook
Piscataquis
Somerset
Penobscot
Washington
Franklin
Hancock
Oxford
Waldo
Kennebec
Androscoggin
Knox
Lincoln
Sagadahoc
Cumberland
York

Maine Wavelet Domain Mapping − Distance

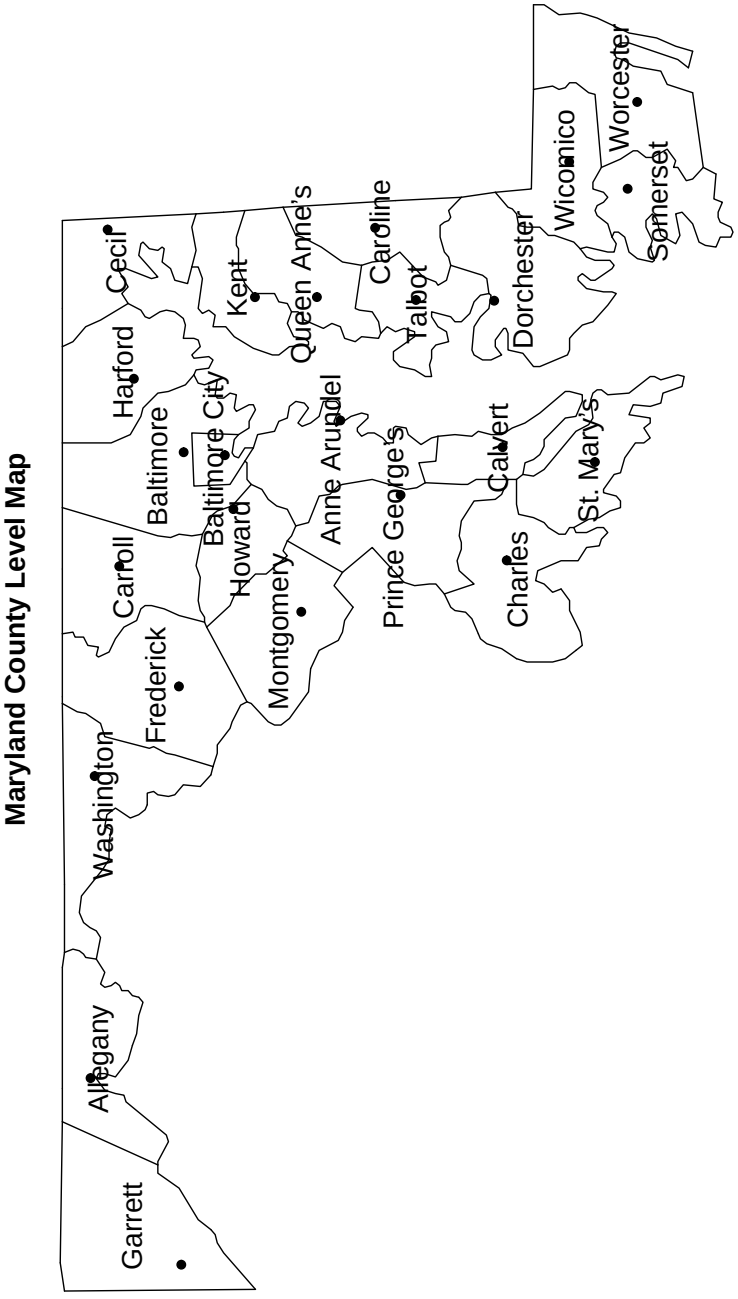
Oxford	Franklin	Somerset	Aroostook
Androscoggin	Kennebec	Piscataquis	Penobscot
Cumberland	Lincoln	Waldo	Hancock
York	Sagadahoc	Knox	Washington

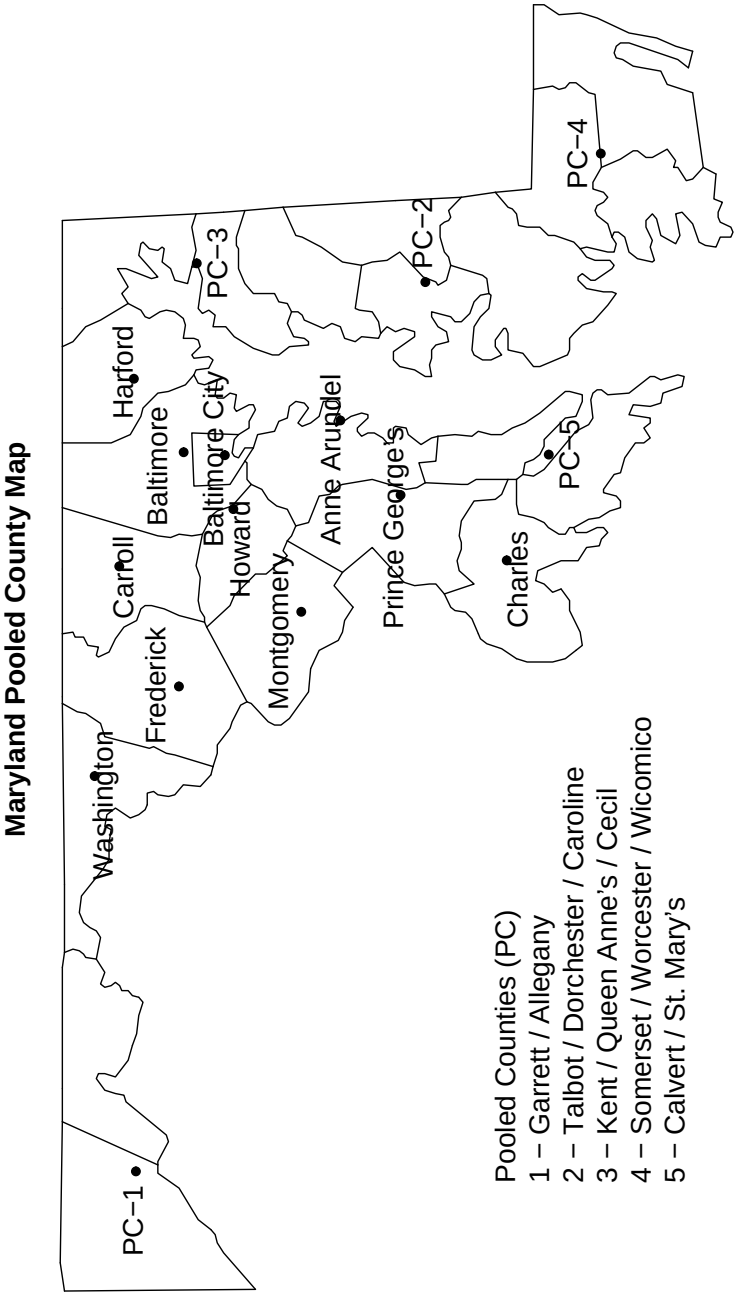
Maine Wavelet Domain Mapping − Direction

Franklin	Somerset	Piscataquis	Aroostook
Oxford	Kennebec	Penobscot	Washington
Androscoggin	Sagadahoc	Waldo	Hancock
York	Cumberland	Lincoln	Knox

Maine Wavelet Domain Mapping − Adjacent

York	Somerset	Piscataquis	Aroostook
Oxford	Franklin	Penobscot	Washington
Androscoggin	Kennebec	Waldo	Hancock
Cumberland	Sagadahoc	Lincoln	Knox


Maryland County Level Map
Allegany
Garrett
Washington
Frederick
Carroll
Baltimore
Harford
Cecil
Baltimore City
Howard
Montgomery
Anne Arundel
Prince George's
Charles
Calvert
St. Mary's
Kent
Queen Anne's
Caroline
Talbot
Dorchester
Wicomico
Somerset
Worcester


Maryland Pooled County Map
PC−1
PC−2
PC−3
PC−4
PC−5
Washington
Frederick
Montgomery
Carroll
Howard
Baltimore
Baltimore City
Harford
Anne Arundel
Prince George's
Charles
Pooled Counties (PC)
1 – Garrett / Allegany
2 – Talbot / Dorchester / Caroline
3 – Kent / Queen Anne's / Cecil
4 – Somerset / Worcester / Wicomico
5 – Calvert / St. Mary's

Maryland Wavelet Domain Mapping – Distance

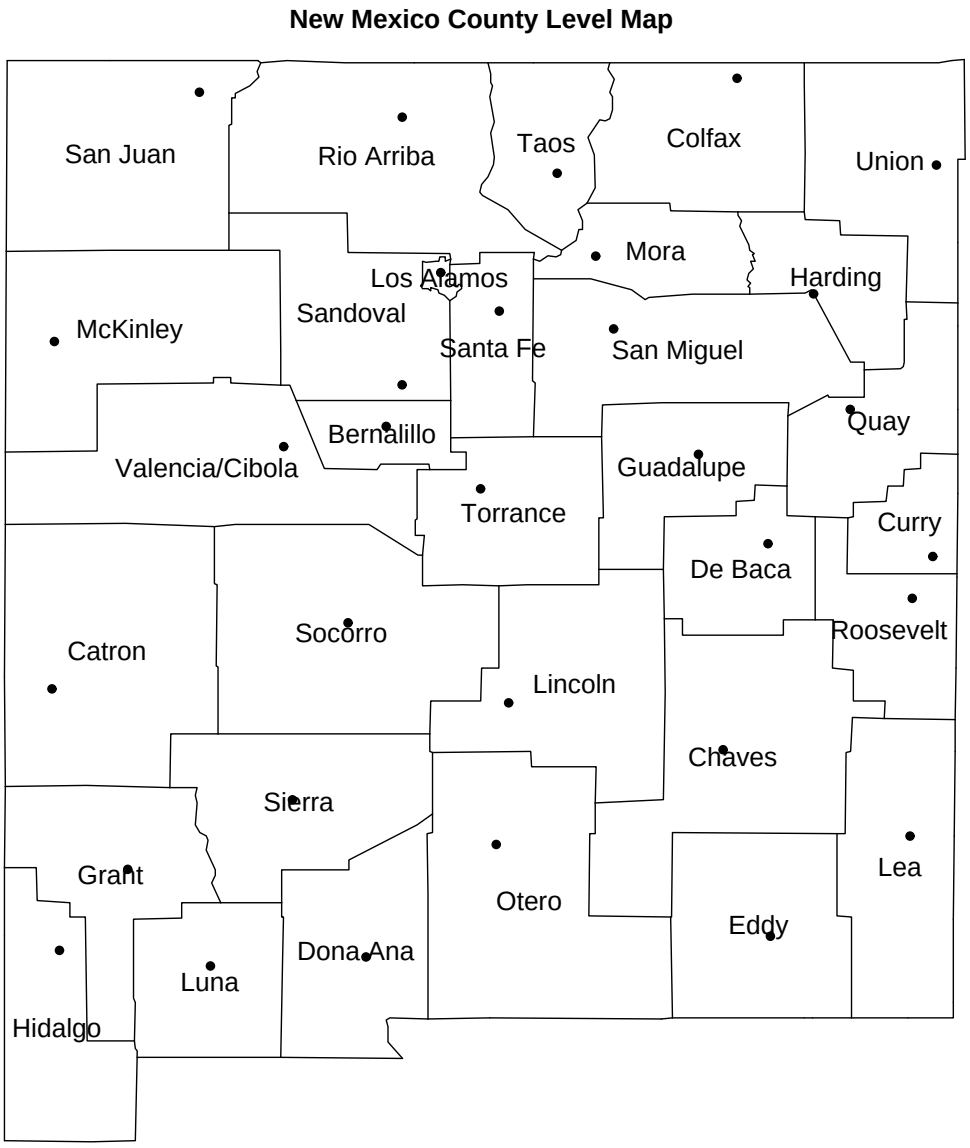
PC–1	Washington	Carroll	Baltimore
Frederick	Howard	Baltimore City	Harford
Montgomery	Prince George's	Anne Arundel	PC–3
Charles	PC–5	PC–2	PC–4

Maryland Wavelet Domain Mapping – Direction

Washington	Carroll	Baltimore	Harford
Frederick	Howard	Baltimore City	PC−3
Montgomery	Prince George's	Anne Arundel	PC−2
PC−1	Charles	PC−5	PC−4

Maryland Wavelet Domain Mapping – Adjacent

PC–1	Washington	Carroll	Frederick
Harford	Baltimore	Howard	Montgomery
PC–3	Baltimore City	Anne Arundel	Prince George's
PC–2	PC–4	PC–5	Charles


New Mexico County Level Map
San Juan
Rio Arriba
Taos
Colfax
Union
Mora
Los Alamos
Harding
Sandoval
McKinley
Santa Fe
San Miguel
Quay
Bernalillo
Valencia/Cibola
Guadalupe
Torrance
Curry
De Baca
Roosevelt
Socorro
Catron
Lincoln
Chaves
Sierra
Grant
Otero
Lea
Eddy
Dona Ana
Luna
Hidalgo

New Mexico Wavelet Domain Mapping − Distance

McKinley	Rio Arriba	Taos	Colfax
Catron	Bernalillo	Santa Fe	Harding
Grant	Socorro	Torrance	De Baca
Hidalgo	Dona Ana	Chaves	Roosevelt
San Juan	Los Alamos	Mora	Union
Valencia/Cibola	Sandoval	San Miguel	Quay
Sierra	Lincoln	Guadalupe	Curry
Luna	Otero	Eddy	Lea

New Mexico Wavelet Domain Mapping − Direction

San Juan	Rio Arriba	Mora	Harding
McKinley	Bernalillo	Santa Fe	Quay
Catron	Socorro	Lincoln	De Baca
Hidalgo	Sierra	Otero	Chaves
Los Alamos	Taos	Colfax	Union
Valencia/Cibola	Sandoval	San Miguel	Curry
Grant	Torrance	Guadalupe	Roosevelt
Luna	Dona Ana	Eddy	Lea

New Mexico Wavelet Domain Mapping − Adjacent

San Juan	Rio Arriba	Taos	McKinley
Sandoval	Los Alamos	Colfax	Union
Bernalillo	Santa Fe	Mora	Harding
Valencia/Cibola	Torrance	Guadalupe	San Miguel
Socorro	Lincoln	De Baca	Quay
Catron	Sierra	Roosevelt	Curry
Grant	Luna	Lea	Eddy
Hidalgo	Dona Ana	Otero	Chaves


California County Level Map
Del Norte
Siskiyou
Modoc
Humboldt
Trinity
Shasta
Lassen
Tehama
Plumas
Glenn
Butte
Sierra
Mendocino
Yuba
Nevada
Colusa
Lake
Sutter
Placer
El Dorado
Yolo
Alpine
Sonoma
Napa
Sacramento
Amador
Solano
Calaveras
Marin
Tuolumne
Mono
Contra Costa
San Joaquin
San Francisco
Alameda
Stanislaus
Mariposa
San Mateo
Santa Clara
Merced
Madera
Santa Cruz
Fresno
San Benito
Inyo
Monterey
Tulare
Kings
San Luis Obispo
Kern
San Bernardino
Santa Barbara
Ventura
Los Angeles
Orange
Riverside
San Diego
Imperial


California Pooled County Map
PC−9
PC−3
Butte
PC−2
PC−8
PC−4
Sonoma
Sacramento
Solano
Marin
Contra Costa
San Joaquin
San Francisco
Alameda
Stanislaus
PC−5
San Mateo
Santa Clara
PC−1
Santa Cruz
Fresno
Monterey
Tulare
PC−6
Kern
San Bernardino
Santa Barbara
Ventura
Los Angeles
Orange
PC−7
San Diego
Pooled Counties (PC)
1 − San Benito / Merced
2 − Nevada / Sutter / Yuba / Sierra / Plumas / Modoc / Lassen
3 − Humboldt / Lake / Mendocino
4 − Napa / Yolo
5 − Madera / Mariposa / Tuolumne / Calaveras / Alpine / Mono / Inyo / Amador
6 − Kings / San Luis Obispo
7 − Imperial / Riverside
8 − El Dorado / Placer
9 − Shasta / Del Norte / Trinity / Siskiyou / Colusa / Glenn / Tehama

California Wavelet Domain Mapping − Distance

PC−9	PC−2	Sacramento	PC−7
PC−4	San Joaquin	PC−5	San Bernardino
San Mateo	PC−1	Tulare	Los Angeles
Alameda	Monterey	Ventura	San Diego
PC−3	Butte	PC−8	Contra Costa
Sonoma	Solano	Stanislaus	Fresno
Marin	Santa Cruz	PC−6	Kern
San Francisco	Santa Clara	Santa Barbara	Orange

California Wavelet Domain Mapping − Direction

PC−3	PC−9	PC−2	PC−8
Sonoma	Solano	San Joaquin	Fresno
San Francisco	Santa Clara	PC−6	San Bernardino
San Mateo	Monterey	Ventura	Los Angeles
PC−4	Butte	Sacramento	PC−5
Marin	Contra Costa	Stanislaus	Tulare
Alameda	PC−1	Kern	PC−7
Santa Cruz	Santa Barbara	San Diego	Orange

California Wavelet Domain Mapping − Adjacent

Santa Cruz	San Mateo	San Francisco	Marin
Santa Clara	Alameda	Contra Costa	Solano
Stanislaus	San Joaquin	Sacramento	Sonoma
PC−1	PC−5	PC−8	PC−3
Fresno	Tulare	PC−2	PC−4
Monterey	PC−6	Butte	PC−9
Santa Barbara	Kern	San Bernardino	PC−7
Ventura	Los Angeles	Orange	San Diego


Louisiana County Level Map
Claiborne
Union
Morehouse
West Carroll
East Carroll
Bossier
Webster
Caddo
Lincoln
Ouachita
Richland
Madison
Bienville
Jackson
Franklin
De Soto
Red River
Caldwell
Tensas
Winn
Natchitoches
La Salle
Catahoula
Sabine
Grant
Concordia
Rapides
Vernon
Avoyelles
West Feliciana
East Feliciana
St. Helena
Washington
Evangeline
Pointe Coupee
Beauregard
Allen
St. Landry
Tangipahoa
East Baton Rouge
West Baton Rouge
Livingston
St. Tammany
Calcasieu
Jefferson Davis
Acadia
Lafayette
St. Martin
Iberville
Ascension
St. John the Baptist
St. James
Orleans
Iberia
Assumption
St. Charles
St. Bernard
Cameron
Vermilion
St. Mary
Jefferson
Lafourche
Terrebonne
Plaquemines

Louisiana Wavelet Domain Mapping – Distance

Bossier	Webster	Bienville	Lincoln	Union	Morehouse	West Carroll	Claiborne
Caddo	Red River	Jackson	Caldwell	Ouachita	Richland	Madison	East Carroll
De Soto	Natchitoches	Winn	La Salle	Catahoula	Franklin	Tensas	Washington
Sabine	Grant	Rapides	Avoyelles	West Feliciana	Concordia	St. Helena	Tangipahoa
Vernon	Allen	Evangeline	St. Landry	Pointe Coupee	East Feliciana	Livingston	St. Tammany
Beauregard	Jefferson Davis	Lafayette	St. Martin	West Baton Rouge	East Baton Rouge	St. John the Baptist	Orleans
Calcasieu	Acadia	Vermilion	Iberia	Iberville	Ascension	St. Charles	St. Bernard
Cameron	Terrebonne	St. Mary	Assumption	St. James	Lafourche	Jefferson	Plaquemines

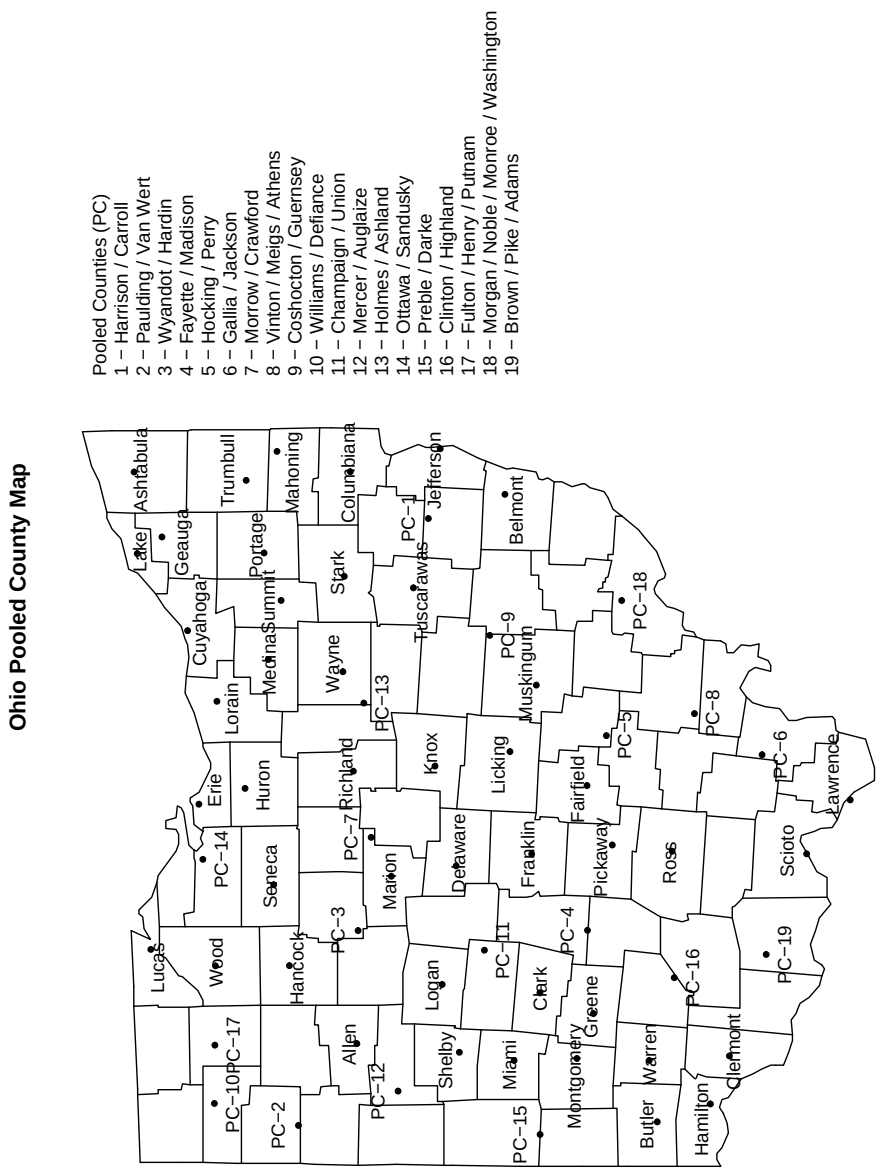
Louisiana Wavelet Domain Mapping – Direction

Webster	Claiborne	Union	Ouachita	Morehouse	Richland	West Carroll	East Carroll
Bossier	Bienville	Jackson	Lincoln	Caldwell	Franklin	Tensas	Madison
Caddo	Red River	Winn	La Salle	Catahoula	Concordia	Tangipahoa	Washington
De Soto	Natchitoches	Grant	Avoyelles	West Feliciana	East Feliciana	St. Helena	St. Tammany
Sabine	Vernon	Rapides	Pointe Coupee	West Baton Rouge	East Baton Rouge	Livingston	St. Bernard
Beauregard	Allen	Evangeline	St. Landry	Iberville	Ascension	St. Charles	Orleans
Calcasieu	Jefferson Davis	Lafayette	St. Martin	Assumption	St. James	St. John the Baptist	Jefferson
Cameron	Acadia	Vermilion	Iberia	St. Mary	Terrebonne	Lafourche	Plaquemines

Louisiana Wavelet Domain Mapping – Adjacent

Morehouse	Red River	Bienville	Claiborne	Union	Pointe Coupee	West Baton Rouge	East Carroll	
West Carroll	Caddo	Bossier	Webster	Lincoln	Concordia	Tensas	Madison	
De Soto	Evangeline	Acadia	Jefferson Davis	Cameron	Catahoula	Franklin	Richland	
Natchitoches	Rapides	Allen	Beauregard	Calcasieu	Ouachita	Caldwell	La Salle	
Sabine	Avoyelles	West Feliciana	St. Helena	Tangipahoa	Jackson	Winn	Grant	
Vernon	St. Landry	East Feliciana	East Baton Rouge	Livingston	St. John the Baptist	St. Charles	St. Bernard	
Lafayette	St. Martin	Iberia	Iberville	Ascension	Lafourche	Jefferson	Plaquemines	
Vermilion	Terrebonne	St. Mary	Assumption	St. James	Washington	St. Tammany	Orleans	


Ohio County Level Map
Williams
Fulton
Lucas
Ottawa
Lake
Ashtabula
Geauga
Cuyahoga
Defiance
Henry
Wood
Sandusky
Erie
Lorain
Trumbull
Paulding
Seneca
Huron
Medina
Summit
Portage
Putnam
Hancock
Mahoning
Van Wert
Allen
Wyandot
Crawford
Ashland
Wayne
Stark
Columbiana
Richland
Hardin
Mercer
Auglaize
Marion
Morrow
Holmes
Carroll
Tuscarawas
Knox
Logan
Jefferson
Shelby
Union
Delaware
Coshocton
Harrison
Champaign
Darke
Miami
Licking
Guernsey
Belmont
Franklin
Muskingum
Clark
Madison
Preble
Montgomery
Greene
Fairfield
Perry
Noble
Monroe
Pickaway
Morgan
Fayette
Hocking
Washington
Butler
Warren
Clinton
Ross
Athens
Vinton
Hamilton
Highland
Pike
Meigs
Clermont
Jackson
Brown
Adams
Scioto
Gallia
Lawrence


Ohio Pooled County Map
Pooled Counties (PC)
1 – Harrison / Carroll
2 – Paulding / Van Wert
3 – Wyandot / Hardin
4 – Fayette / Madison
5 – Hocking / Perry
6 – Gallia / Jackson
7 – Morrow / Crawford
8 – Vinton / Meigs / Athens
9 – Coshocton / Guernsey
10 – Williams / Defiance
11 – Champaign / Union
12 – Mercer / Auglaize
13 – Holmes / Ashland
14 – Ottawa / Sandusky
15 – Preble / Darke
16 – Clinton / Highland
17 – Fulton / Henry / Putnam
18 – Morgan / Noble / Monroe / Washington
19 – Brown / Pike / Adams
Ashtabula
Trumbull
Mahoning
Columbiana
Lake
Geauga
Portage
Jefferson
PC−1
Belmont
Cuyahoga
Summit
Stark
Tuscarawas
PC−18
Medina
Wayne
PC−9
Muskingum
Lorain
PC−13
PC−8
Erie
Huron
Richland
Knox
Licking
PC−5
PC−6
Lawrence
PC−14
Seneca
PC−7
Delaware
Fairfield
Franklin
Pickaway
Ross
Scioto
Marion
PC−3
PC−4
Lucas
Wood
Hancock
Logan
PC−11
Clark
Greene
PC−16
PC−19
PC−17
Allen
Shelby
Miami
Montgomery
Warren
Clermont
PC−10
PC−12
PC−2
PC−15
Butler
Hamilton

Ohio Wavelet Domain Mapping – Distance

PC−2	PC−10	PC−17	Wood	Lucas	Mahoning	Trumbull	Ashtabula
PC−15	PC−12	Allen	Hancock	PC−14	Erie	Geauga	Lake
Miami	Shelby	Logan	PC−3	Seneca	Huron	Lorain	Cuyahoga
Montgomery	Clark	PC−11	Marion	PC−7	Richland	Medina	Summit
Warren	Greene	PC−4	Delaware	Knox	PC−13	Wayne	Portage
Clermont	PC−16	Pickaway	Franklin	Licking	Muskingum	Tuscarawas	Stark
Butler	PC−19	Ross	Fairfield	PC−5	PC−9	PC−1	Columbiana
Hamilton	Lawrence	Scioto	PC−6	PC−8	PC−18	Belmont	Jefferson

Ohio Wavelet Domain Mapping – Direction

PC−10	Wood	Lucas	PC−14	Erie	Lorain	Lake	Ashtabula
PC−2	PC−17	Hancock	Seneca	Huron	Cuyahoga	Geauga	Portage
PC−12	Allen	PC−3	PC−7	Richland	Medina	Summit	Trumbull
Miami	Shelby	Logan	Marion	PC−13	Wayne	Stark	Mahoning
PC−15	Clark	PC−11	Delaware	Knox	Licking	Tuscarawas	Columbiana
Montgomery	Greene	PC−4	Franklin	Pickaway	Muskingum	PC−1	Jefferson
Butler	Warren	PC−16	Ross	Fairfield	PC−5	PC−9	Belmont
Hamilton	Clermont	PC−19	Scioto	Lawrence	PC−6	PC−8	PC−18

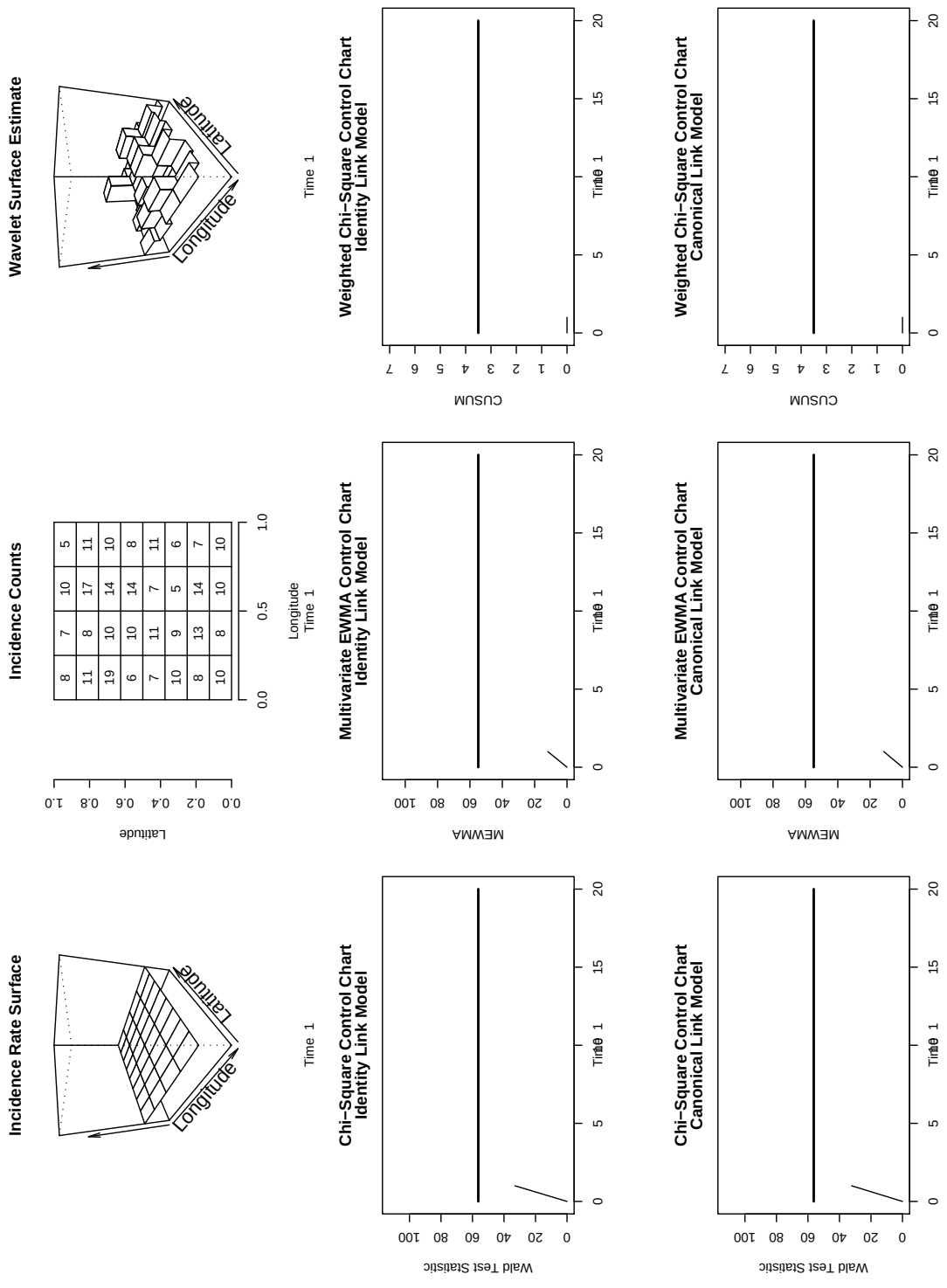
Ohio Wavelet Domain Mapping – Adjacent

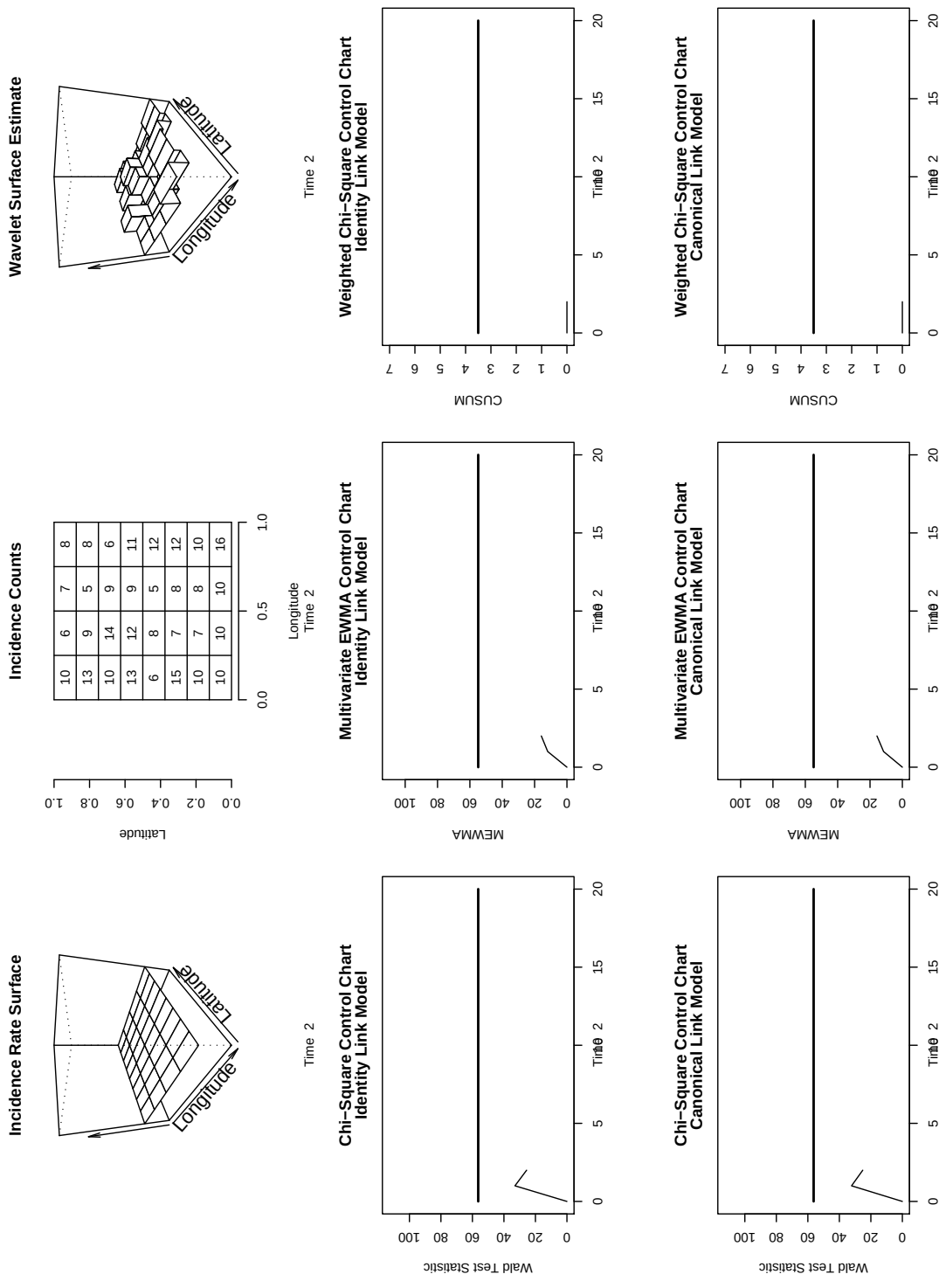
Lawrence	Lake	Cuyahoga	Geauga	Ashtabula	Trumbull	Mahoning	Columbiana
PC−2	PC−10	Summit	Portage	PC−9	Muskingum	Stark	PC−1
Allen	PC−17	Jefferson	Belmont	PC−18	PC−5	Clark	Tuscarawas
Hancock	Wood	Seneca	PC−7	PC−8	Ross	PC−4	PC−16
PC−19	Lucas	PC−14	Clermont	Hamilton	Pickaway	Greene	Warren
PC−6	Scioto	Erie	Knox	Fairfield	Franklin	Montgomery	Butler
Huron	Lorain	Medina	Delaware	Licking	PC−11	Miami	PC−15
Richland	PC−13	Wayne	Marion	PC−3	Logan	Shelby	PC−12

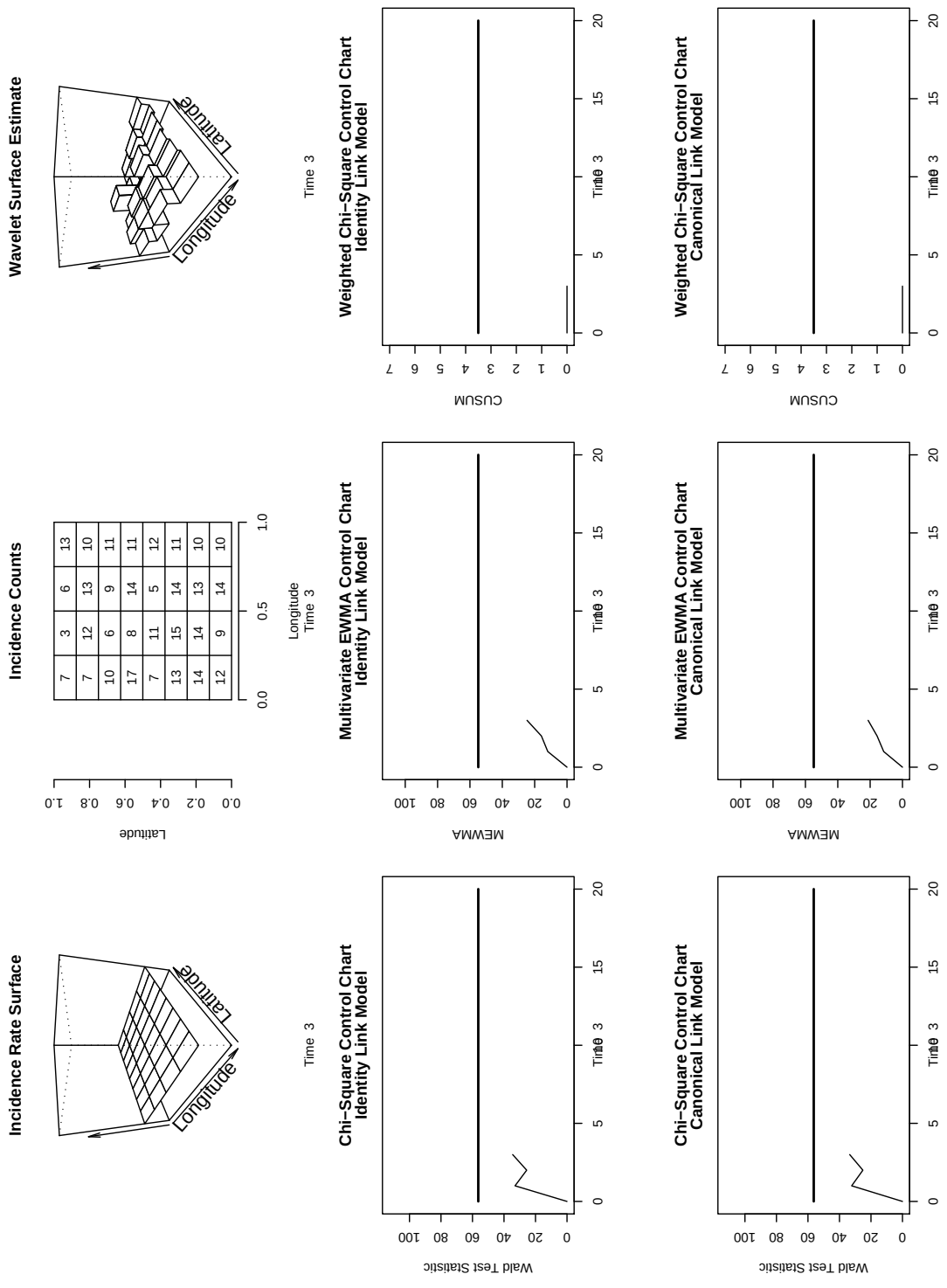
Appendix D

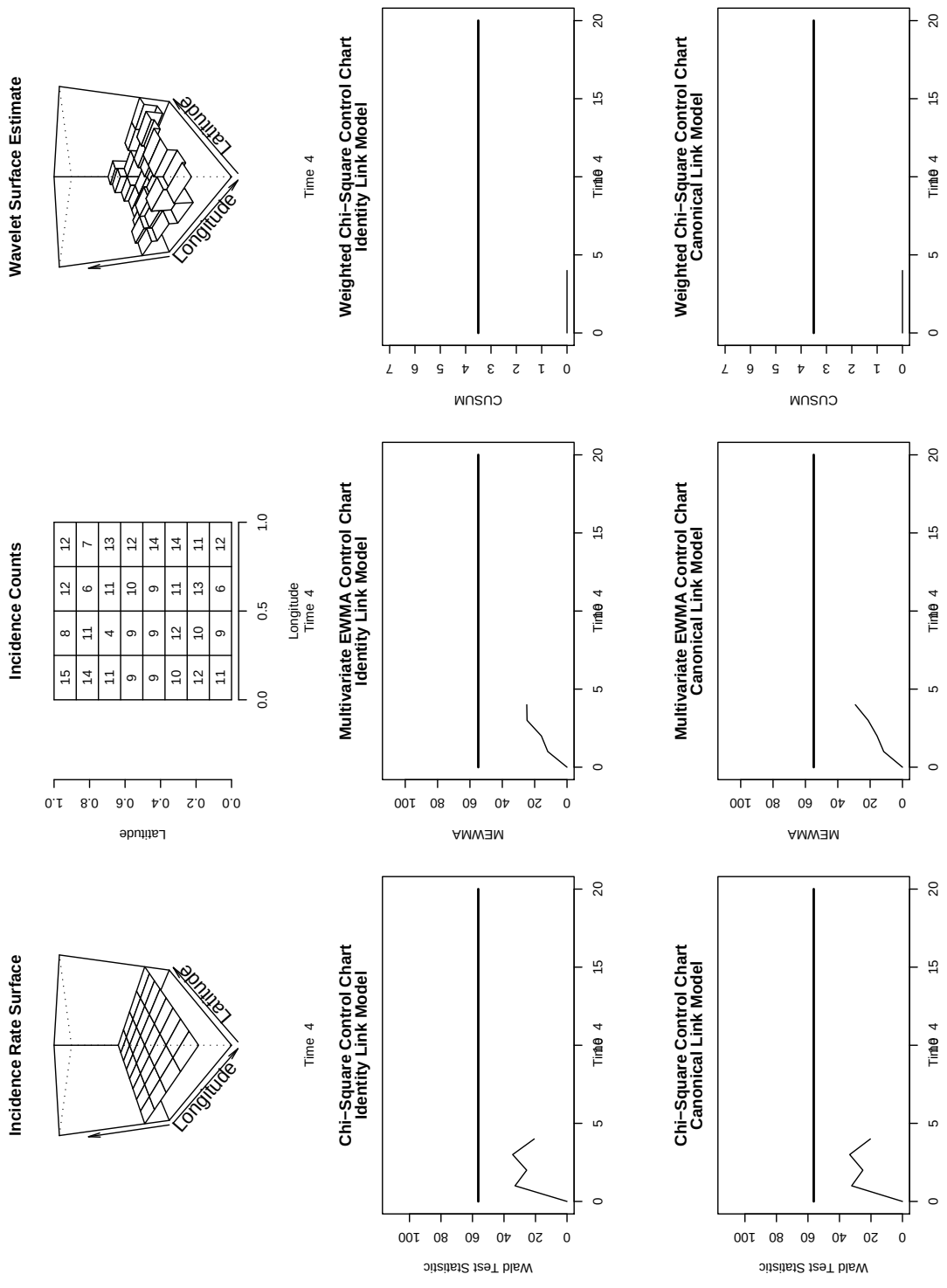
Full Demonstration of the Wavelet-Based Disease Surveillance Method

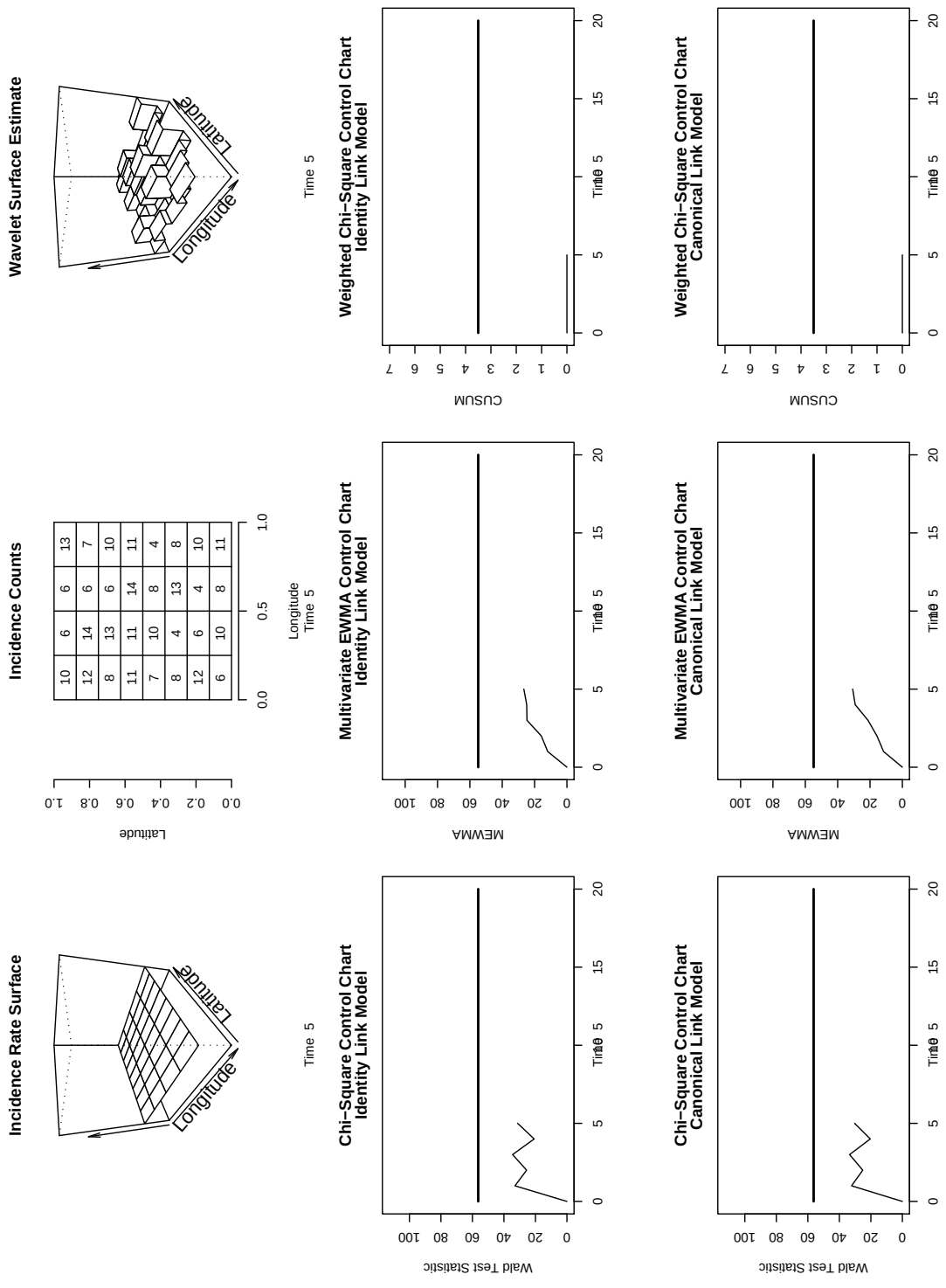
To show how the wavelet-based disease surveillance method works prospectively, the first demonstration of this method discussed in Chapter 3, Section 3.4 is shown in this Appendix. This demonstration showed the effect of a disease cluster in the northern part of a geographical region containing 32 subregions on the chi-square, MEWMA, and Weighted χ^2 control charts using both the identity and canonical link models. For each time interval in this demonstration, the incidence rate surface, the observation generated from that surface, and the estimated surface are given. The control charts for monitoring the incidence rate surface are also shown for each time interval. These charts show the statistic plotted for the current time interval and all previous time intervals.

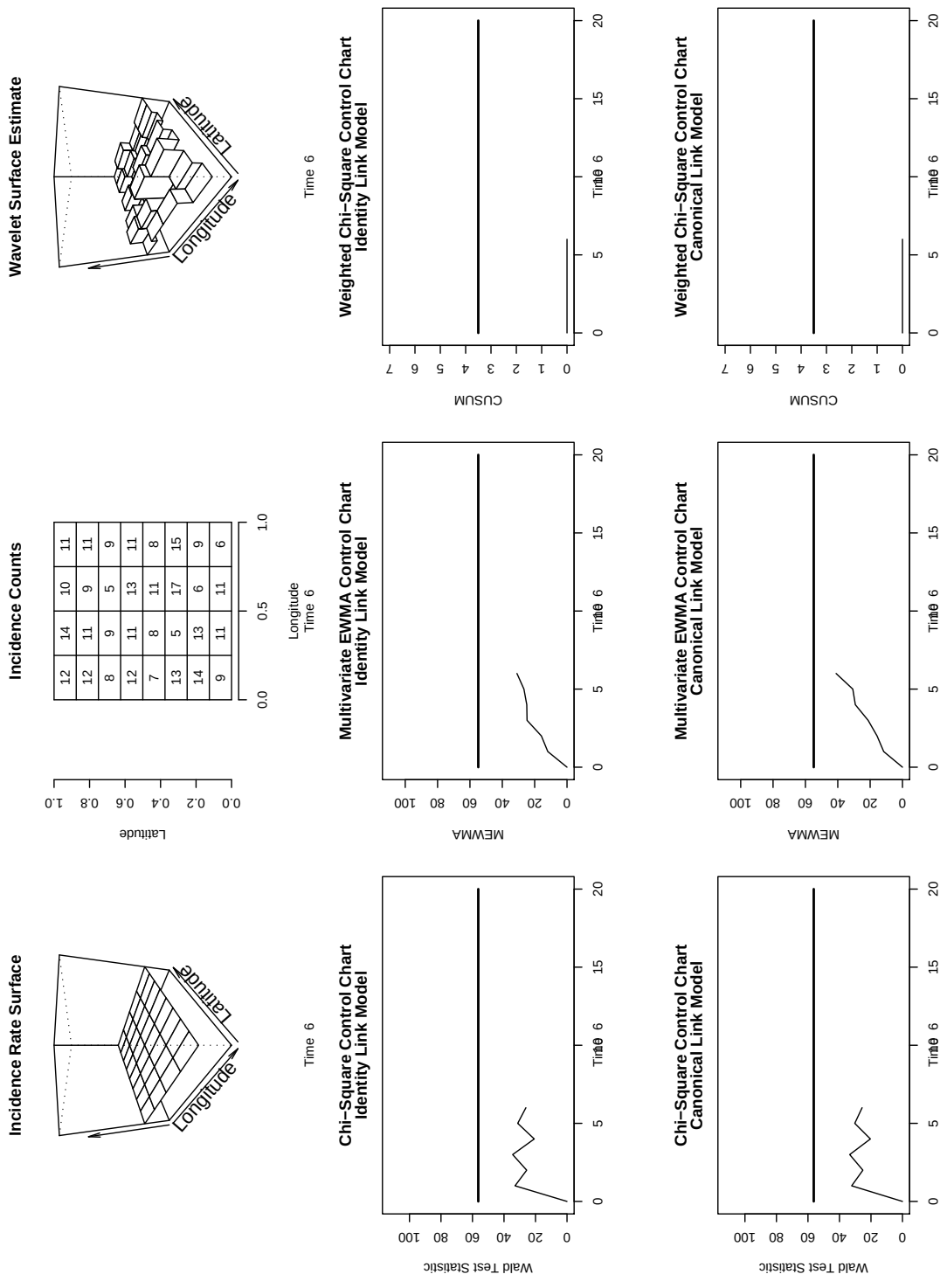

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 1
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

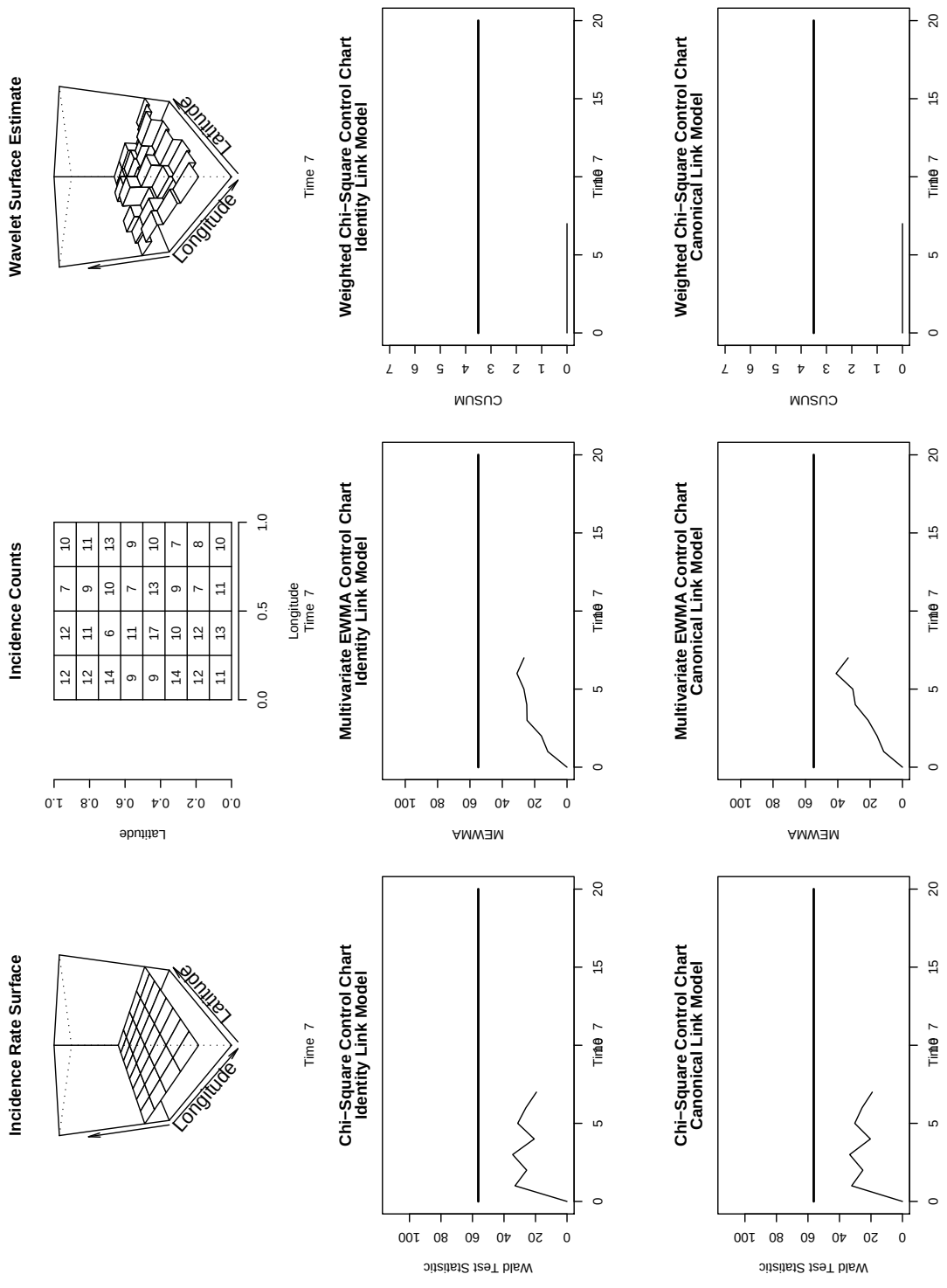

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 2
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time 2
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time 2
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time 2
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time 2
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time 2
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time 2


Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 3
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

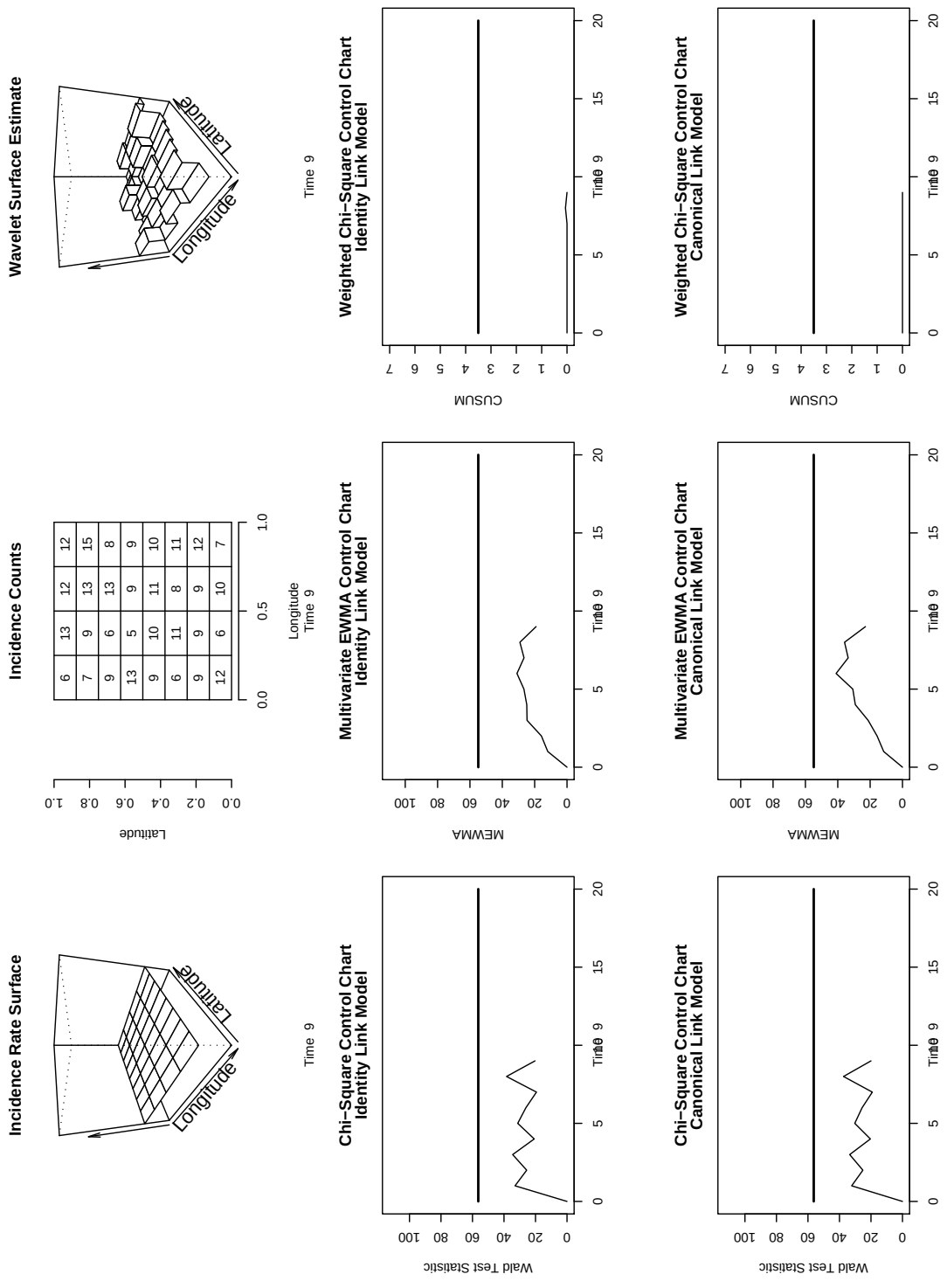

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 4
Wavelet Surface Estimate
Latitude
Longitude
Time 4
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

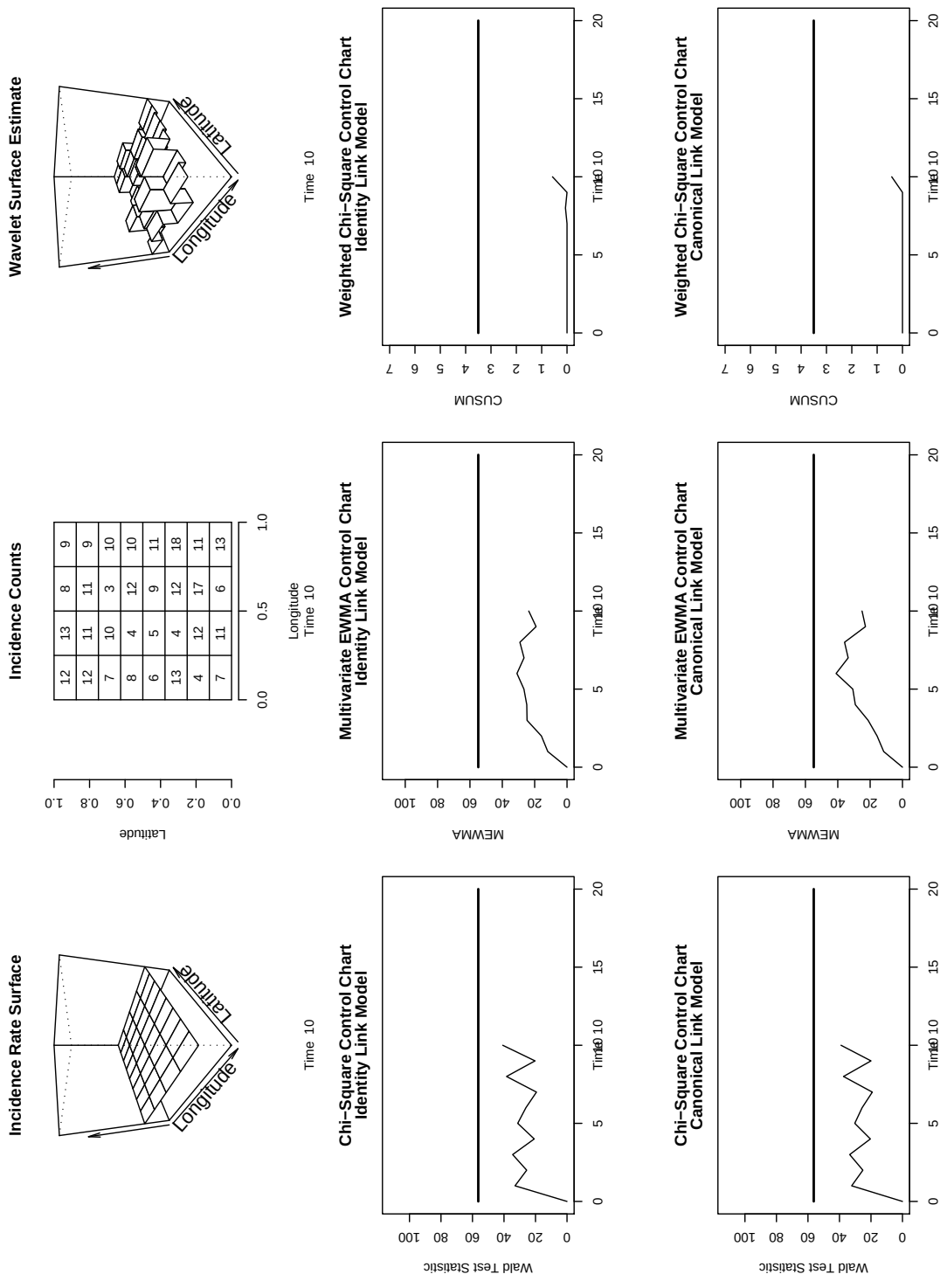

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 5
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

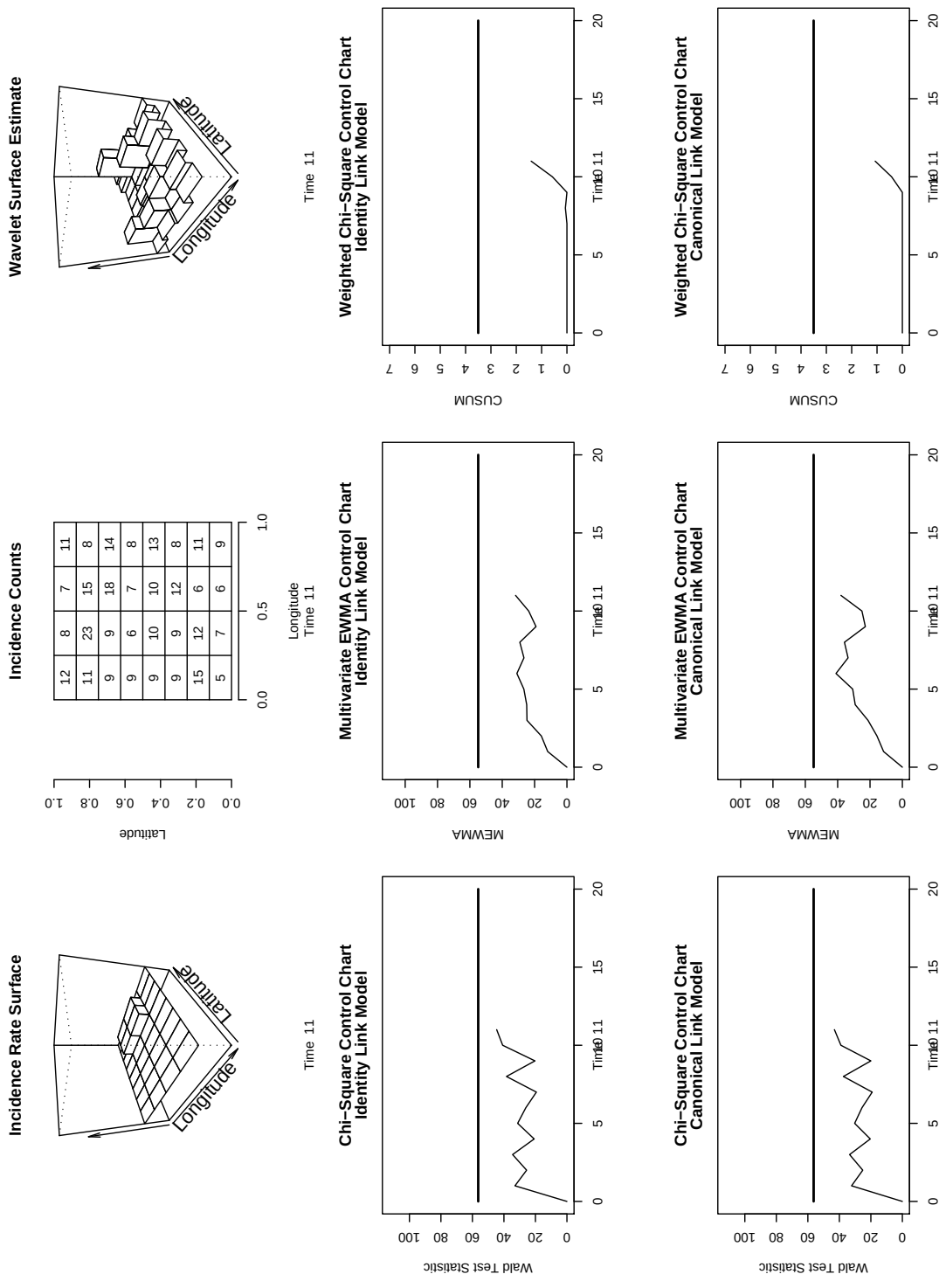

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart Identity Link Model
Wald Test Statistic
Time 6
Multivariate EWMA Control Chart Identity Link Model
MEWMA
Time 6
Weighted Chi−Square Control Chart Identity Link Model
CUSUM
Time 6
Chi−Square Control Chart Canonical Link Model
Wald Test Statistic
Multivariate EWMA Control Chart Canonical Link Model
MEWMA
Weighted Chi−Square Control Chart Canonical Link Model
CUSUM

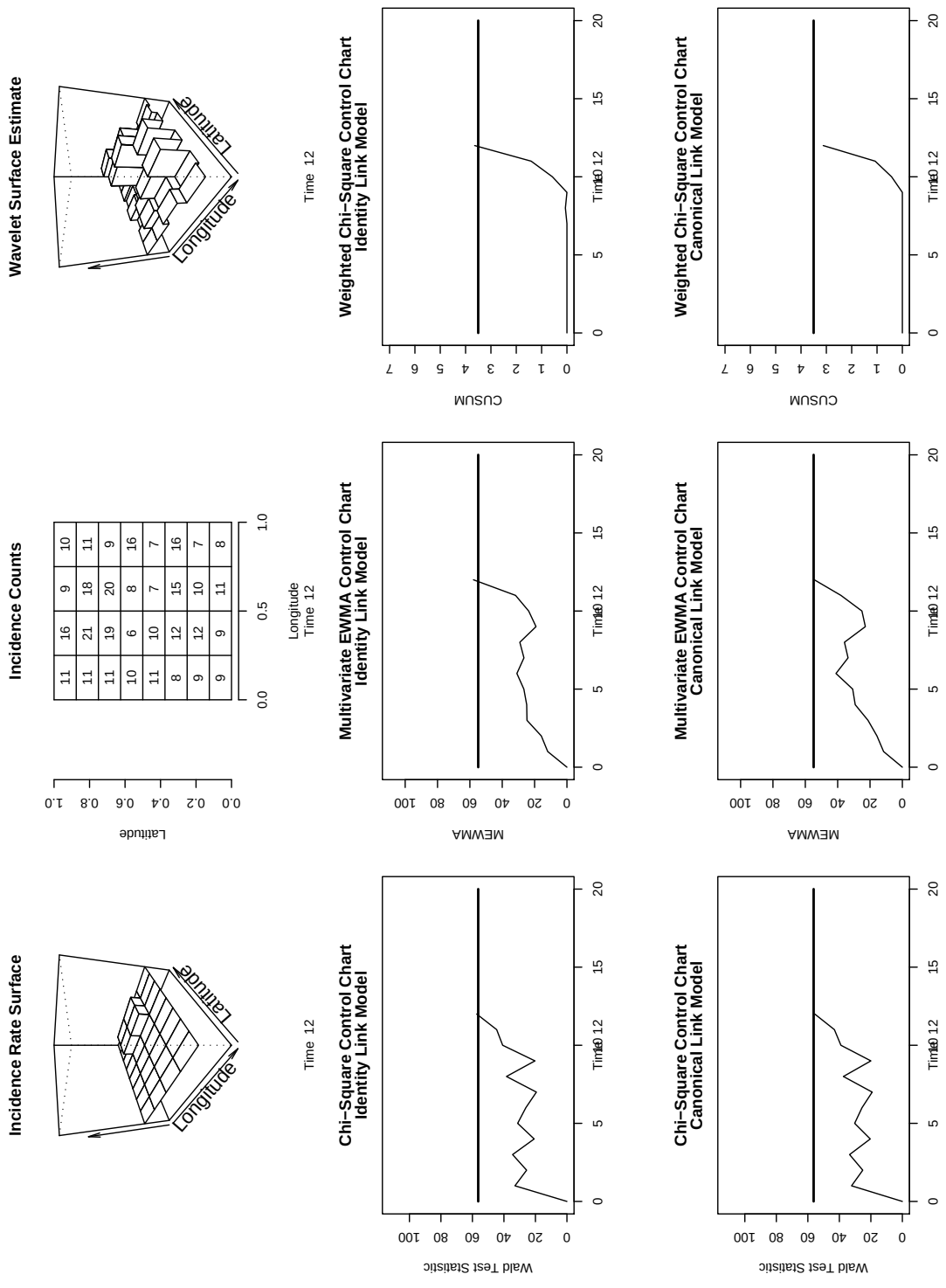

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 7
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

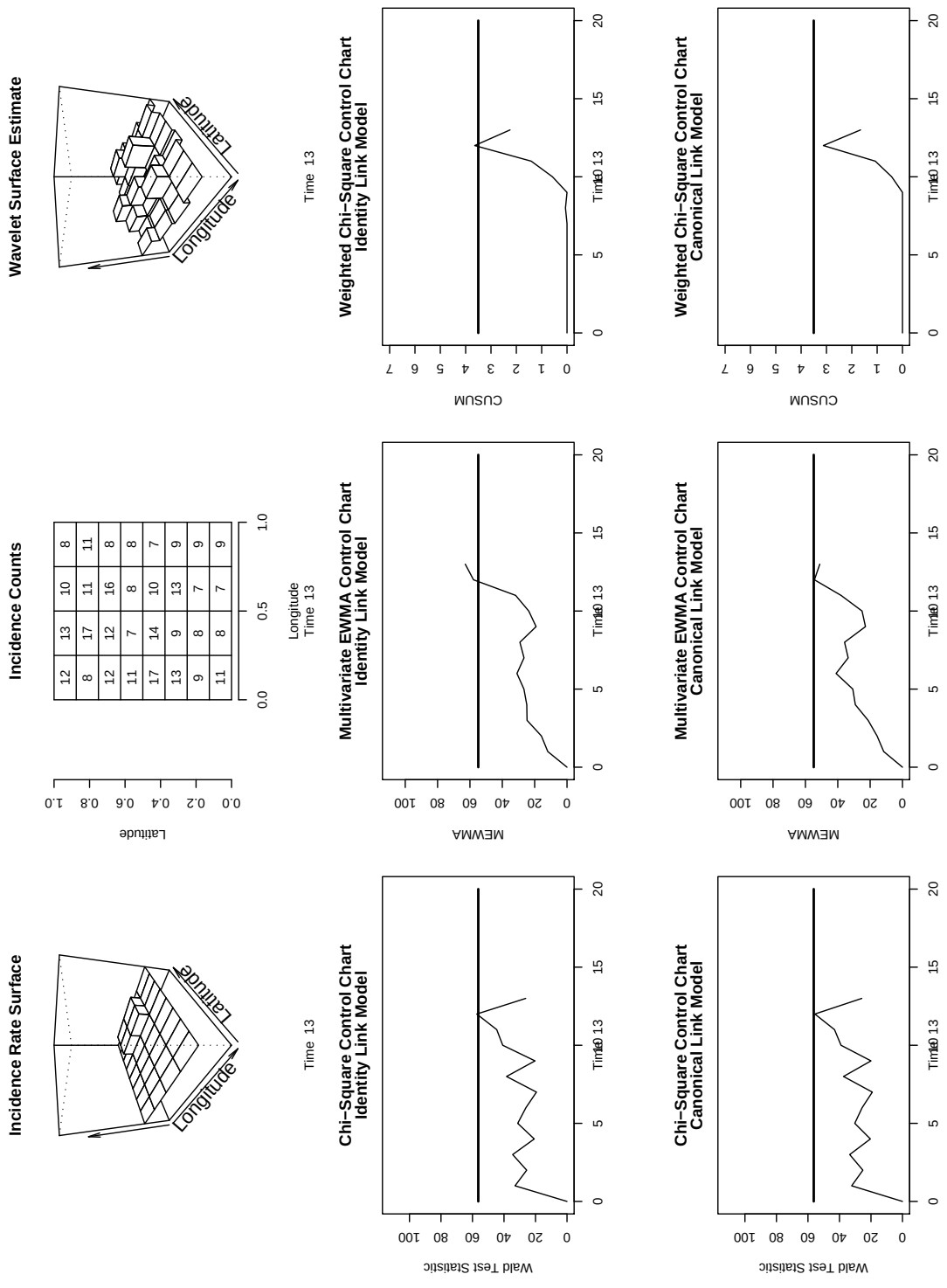
Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 8
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time 8
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time 8
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time 8
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time 8
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time 8
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time 8

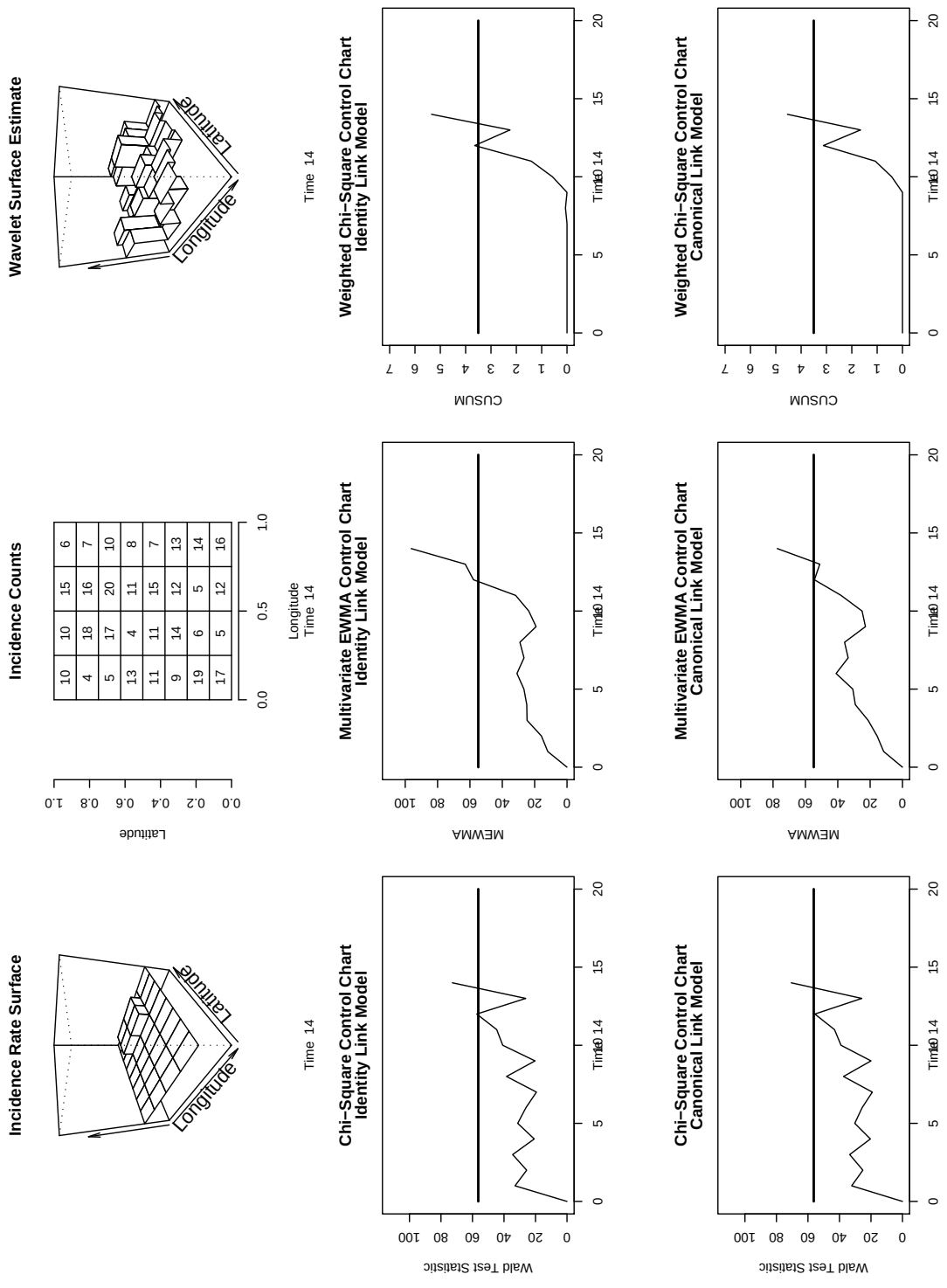

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 9
Wavelet Surface Estimate
Latitude
Longitude
Time 9
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

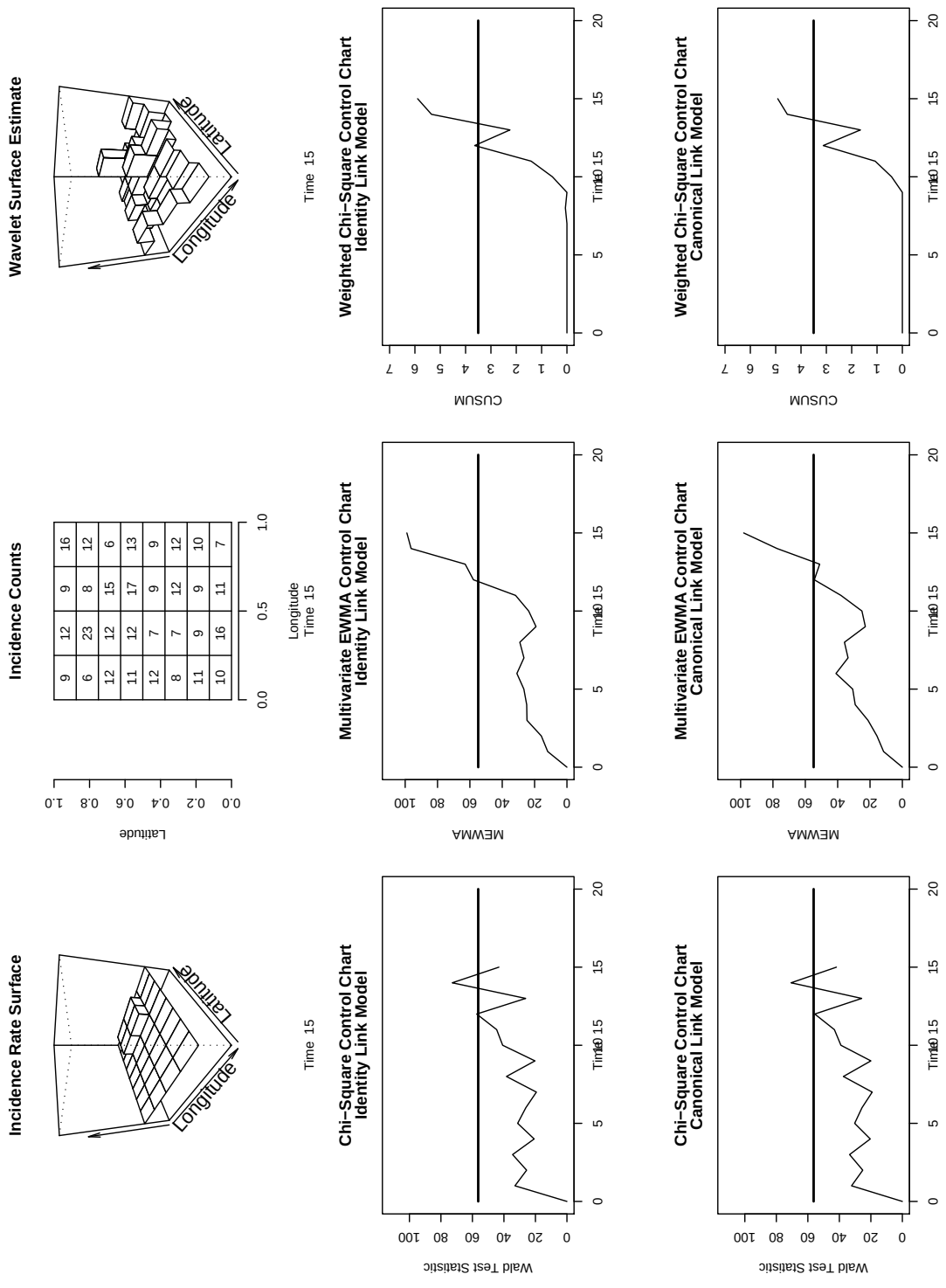

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 10
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

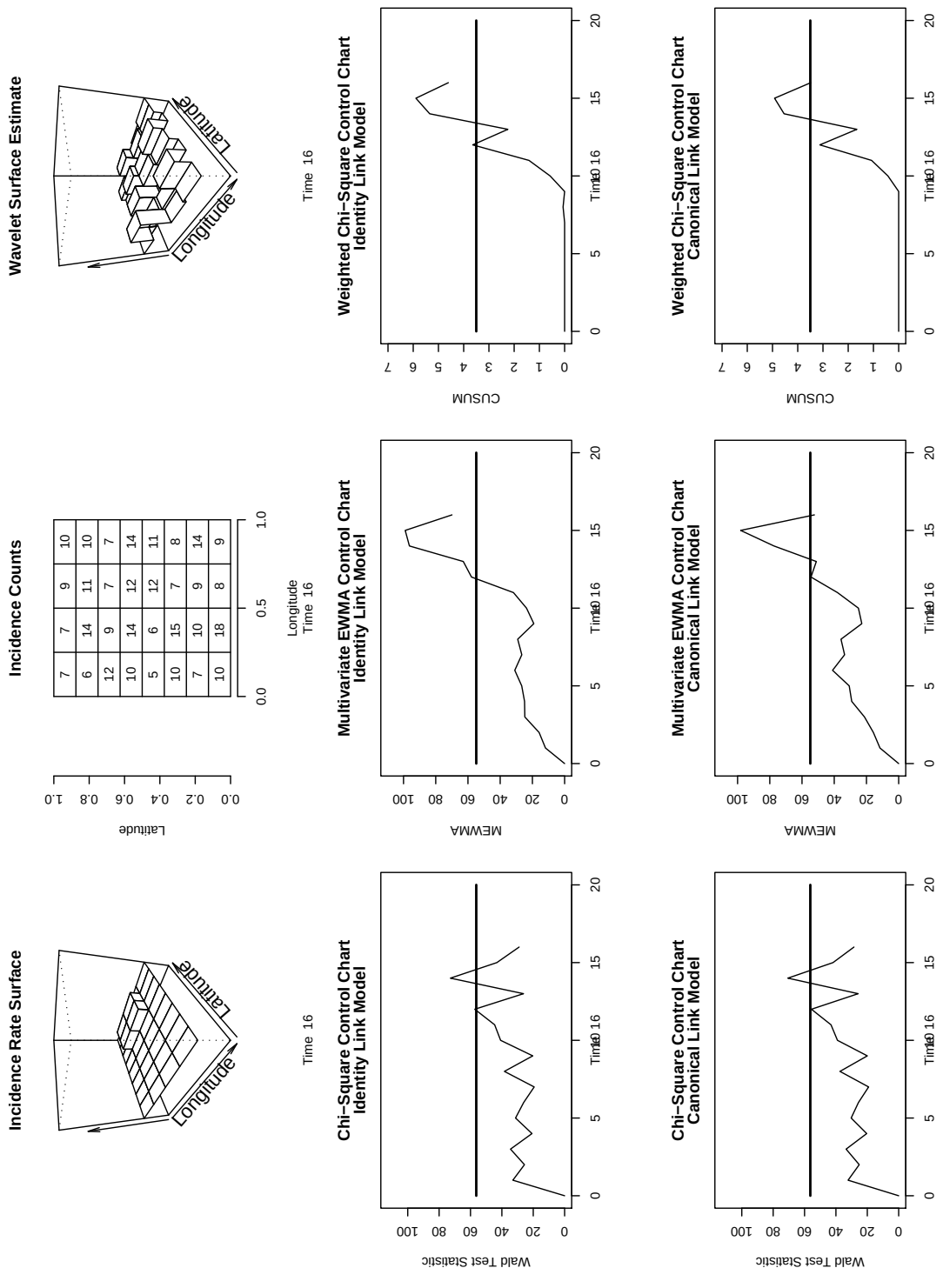

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 11
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

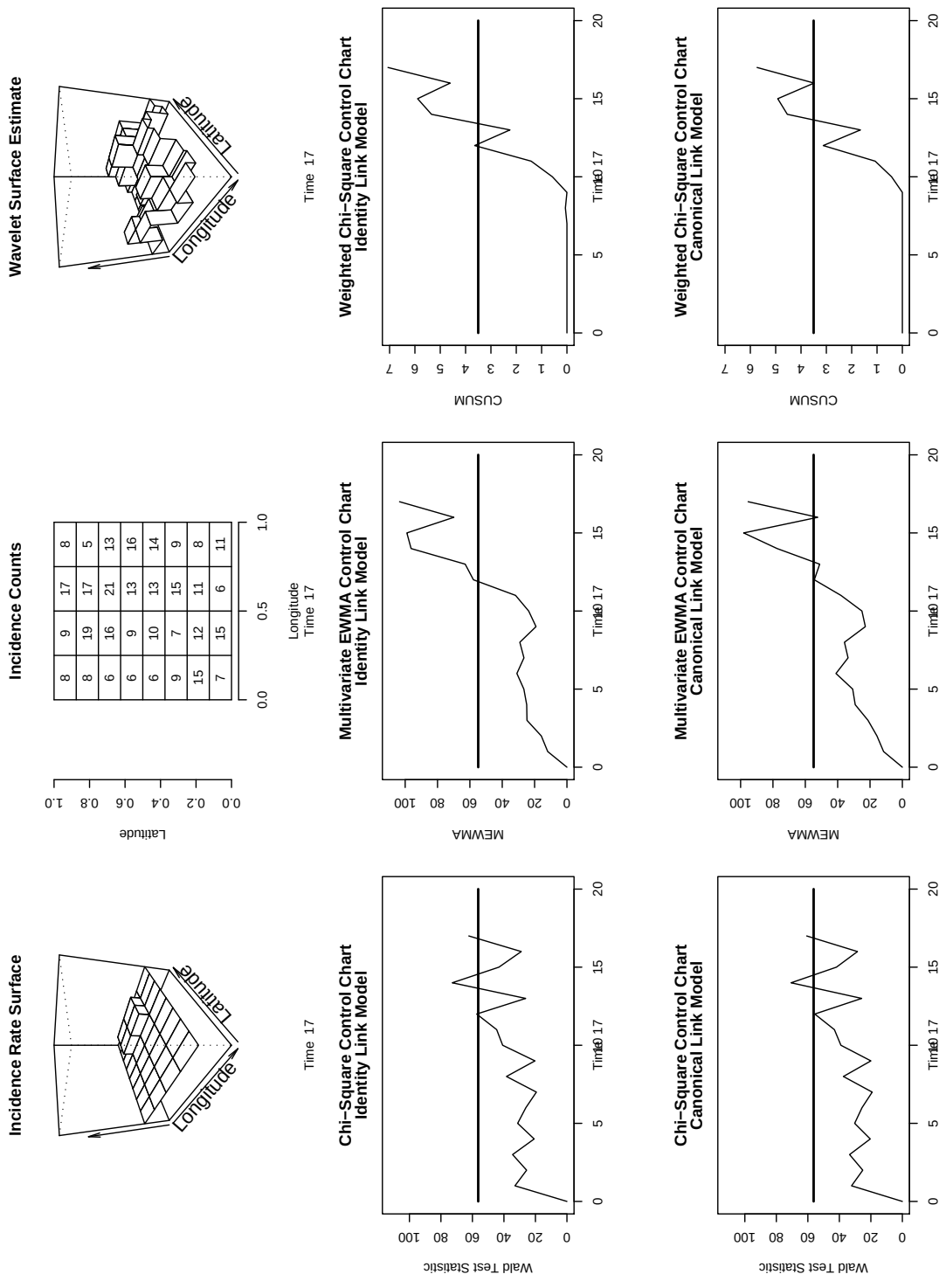

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

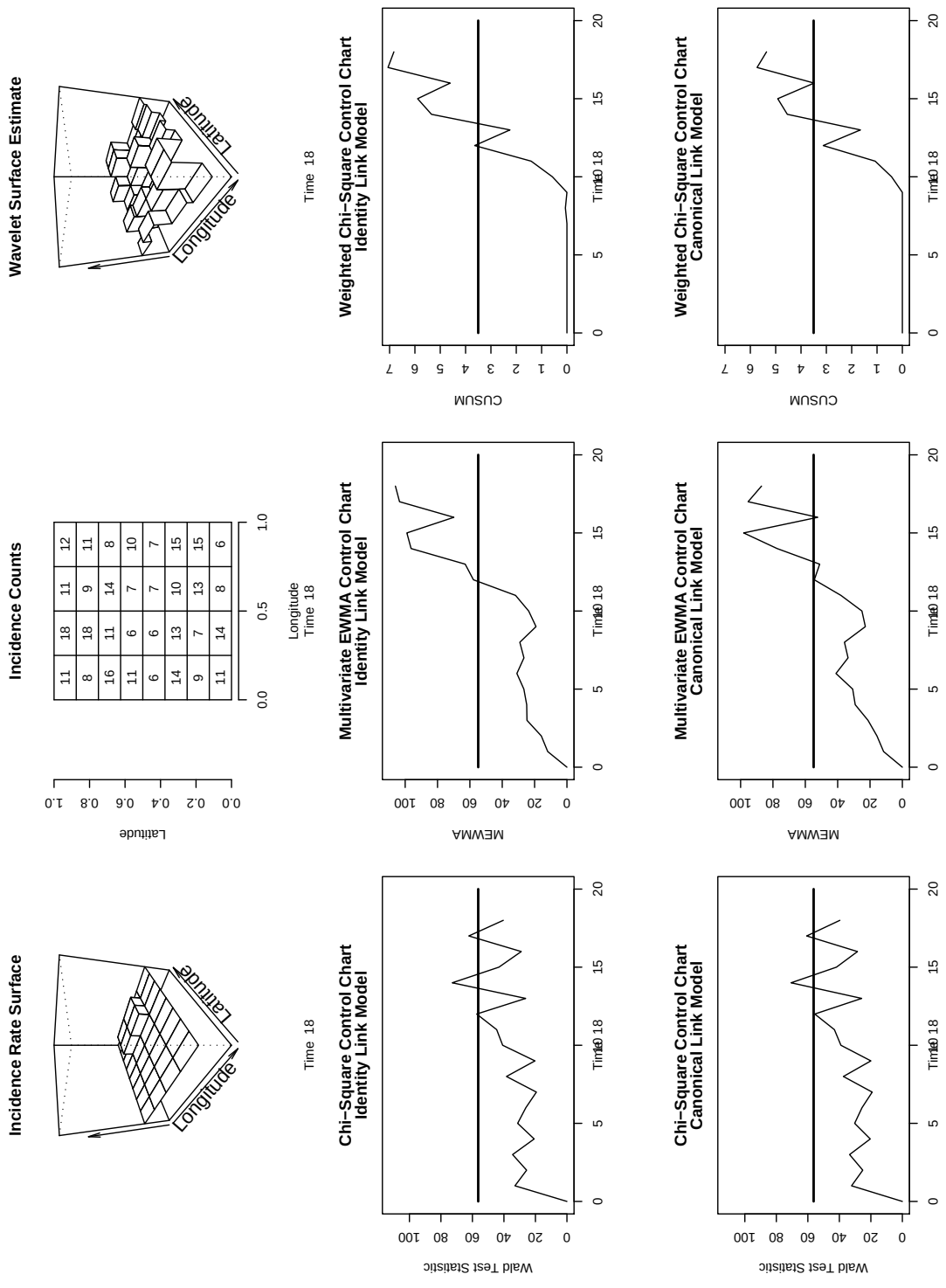

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 13
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

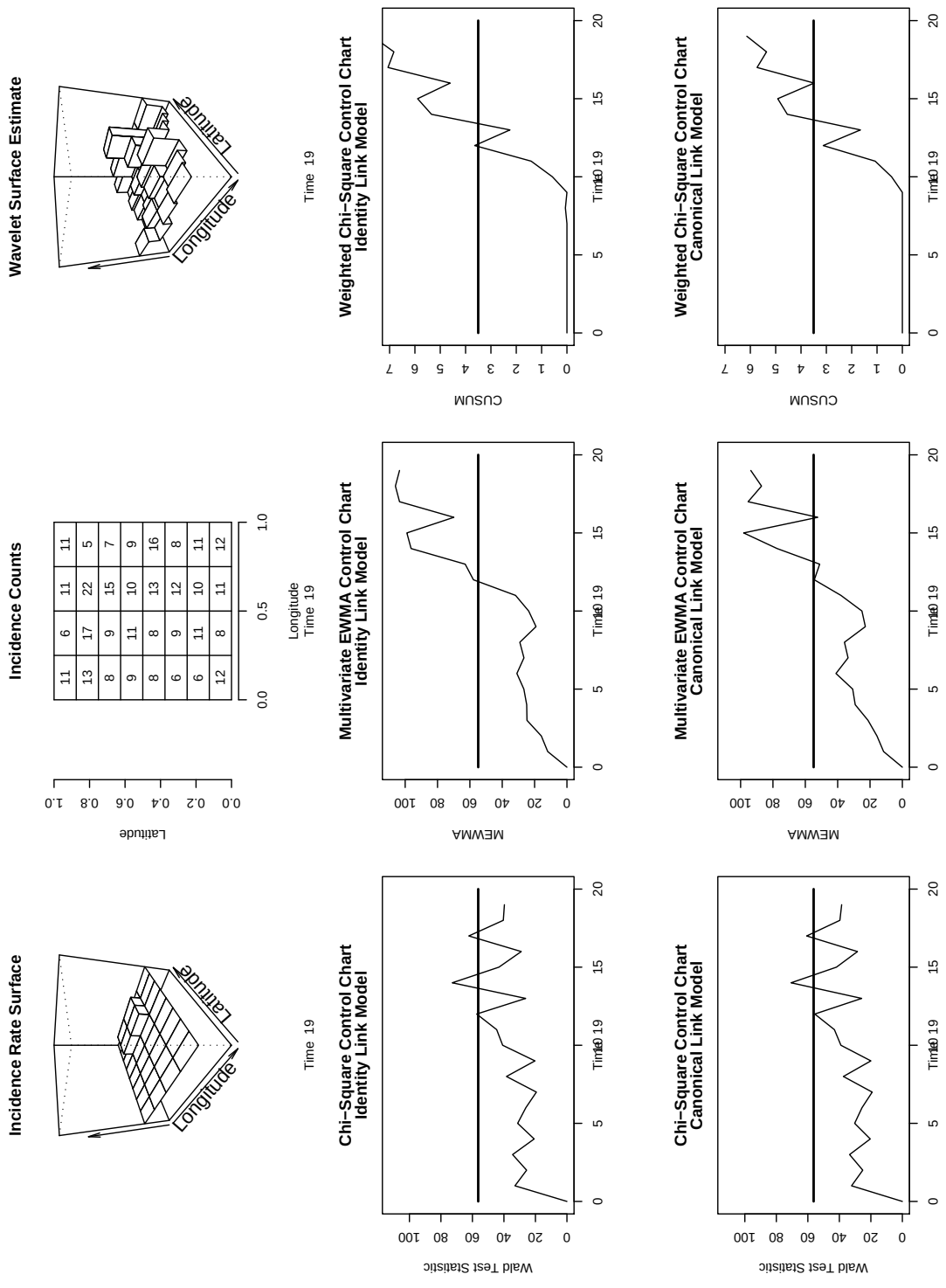

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time 14
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Longitude
Time 14
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time 14
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM

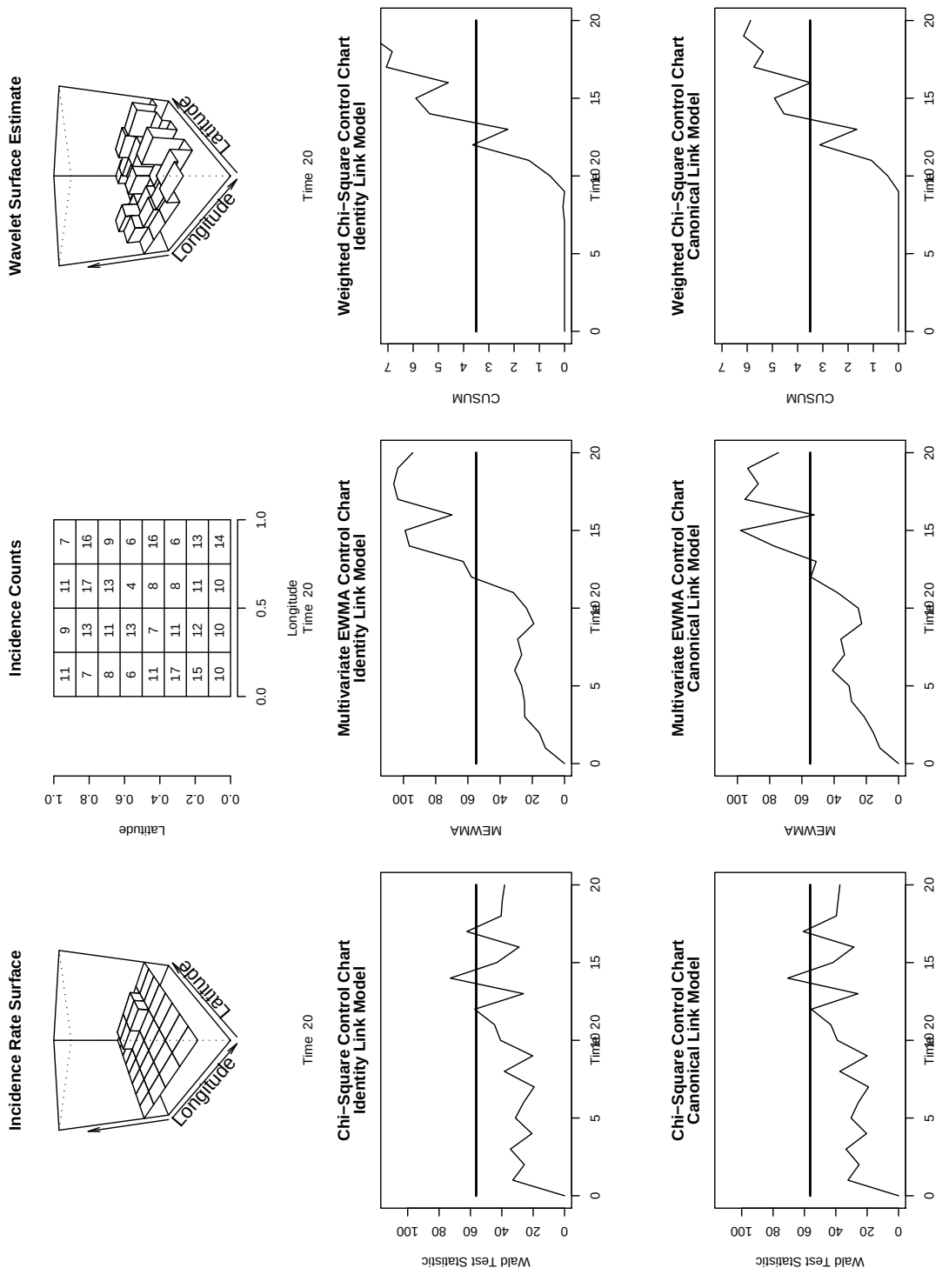

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time 15
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time 15
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time 15
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM


Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 16
Wavelet Surface Estimate
Latitude
Longitude
Time 16
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time


Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 17
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time


Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Wavelet Surface Estimate
Chi−Square Control Chart
Identity Link Model
Multivariate EWMA Control Chart
Identity Link Model
Weighted Chi−Square Control Chart
Identity Link Model
Chi−Square Control Chart
Canonical Link Model
Multivariate EWMA Control Chart
Canonical Link Model
Weighted Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
MEWMA
CUSUM
Time 18

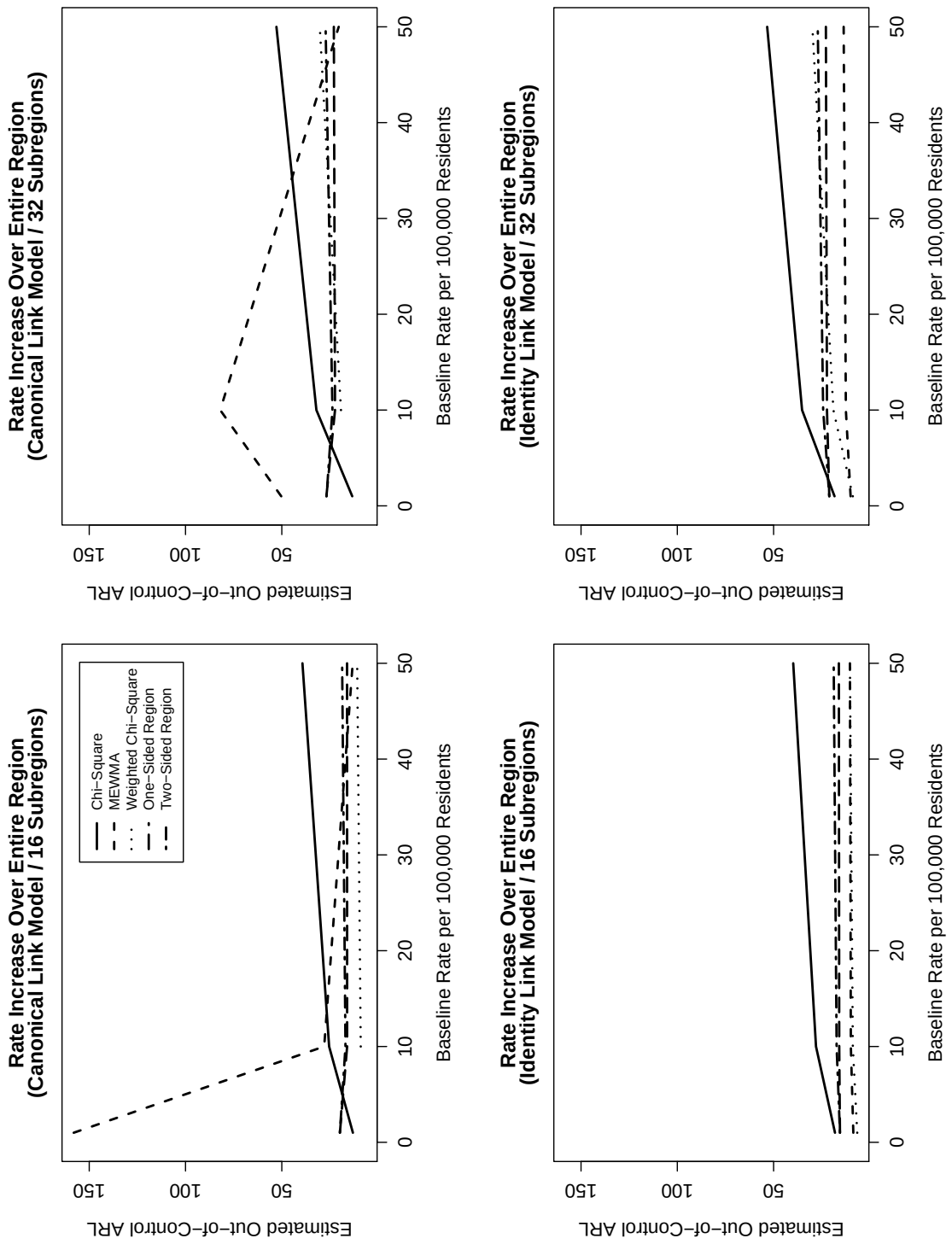

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 19
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

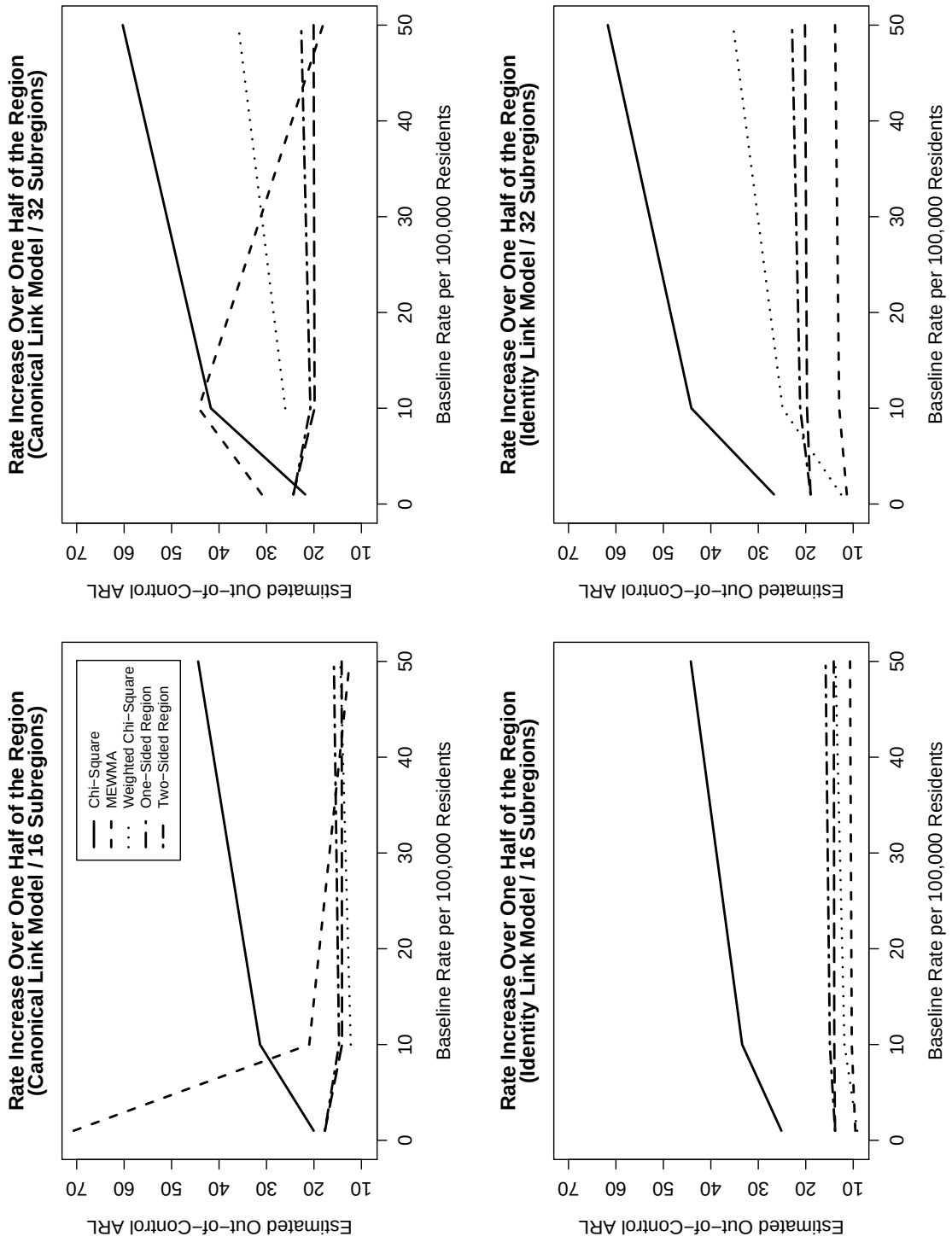

Incidence Rate Surface
Latitude
Longitude
Incidence Counts
Latitude
Longitude
Time 20
Wavelet Surface Estimate
Latitude
Longitude
Chi−Square Control Chart
Identity Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Identity Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Identity Link Model
CUSUM
Time
Chi−Square Control Chart
Canonical Link Model
Wald Test Statistic
Time
Multivariate EWMA Control Chart
Canonical Link Model
MEWMA
Time
Weighted Chi−Square Control Chart
Canonical Link Model
CUSUM
Time

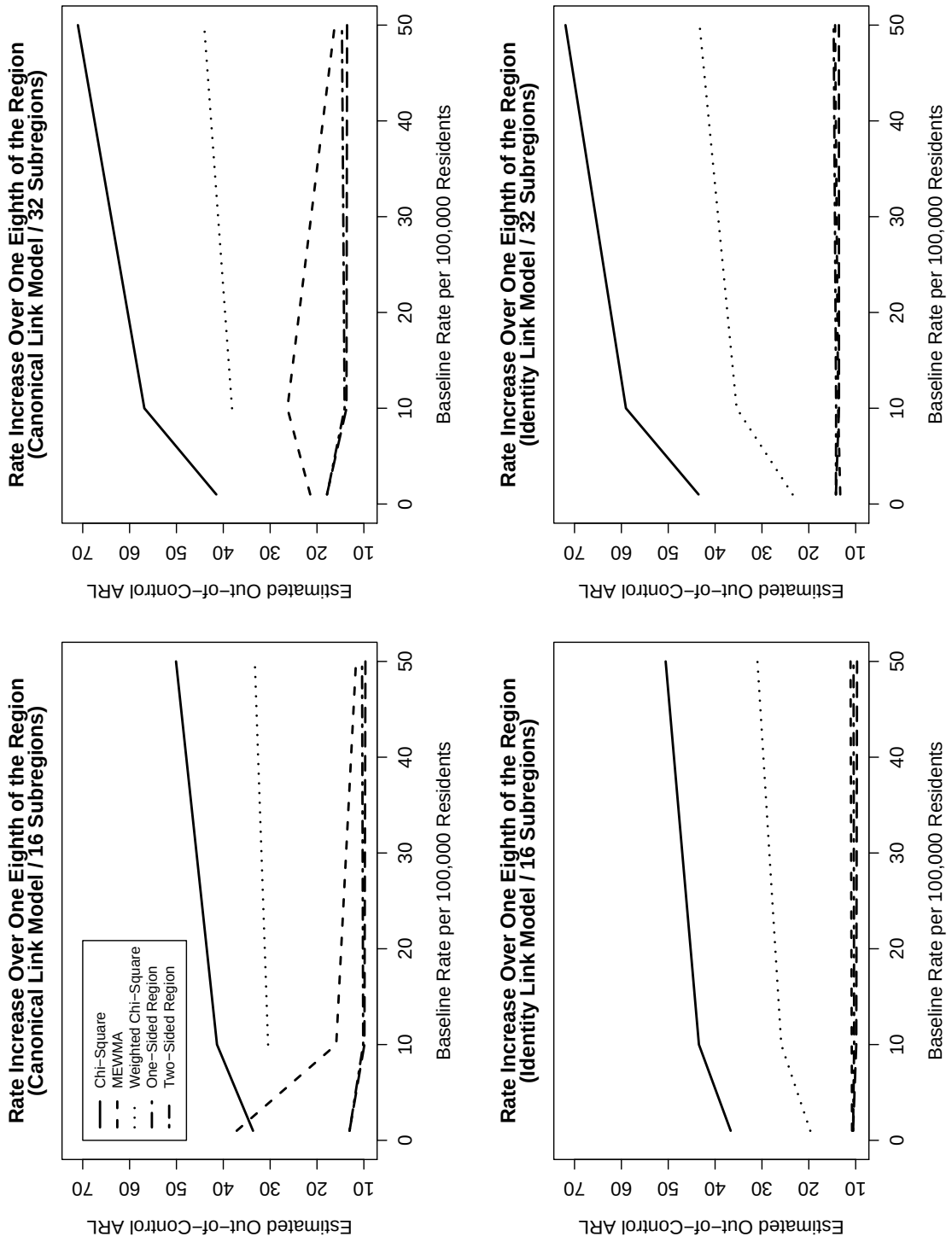
Appendix E

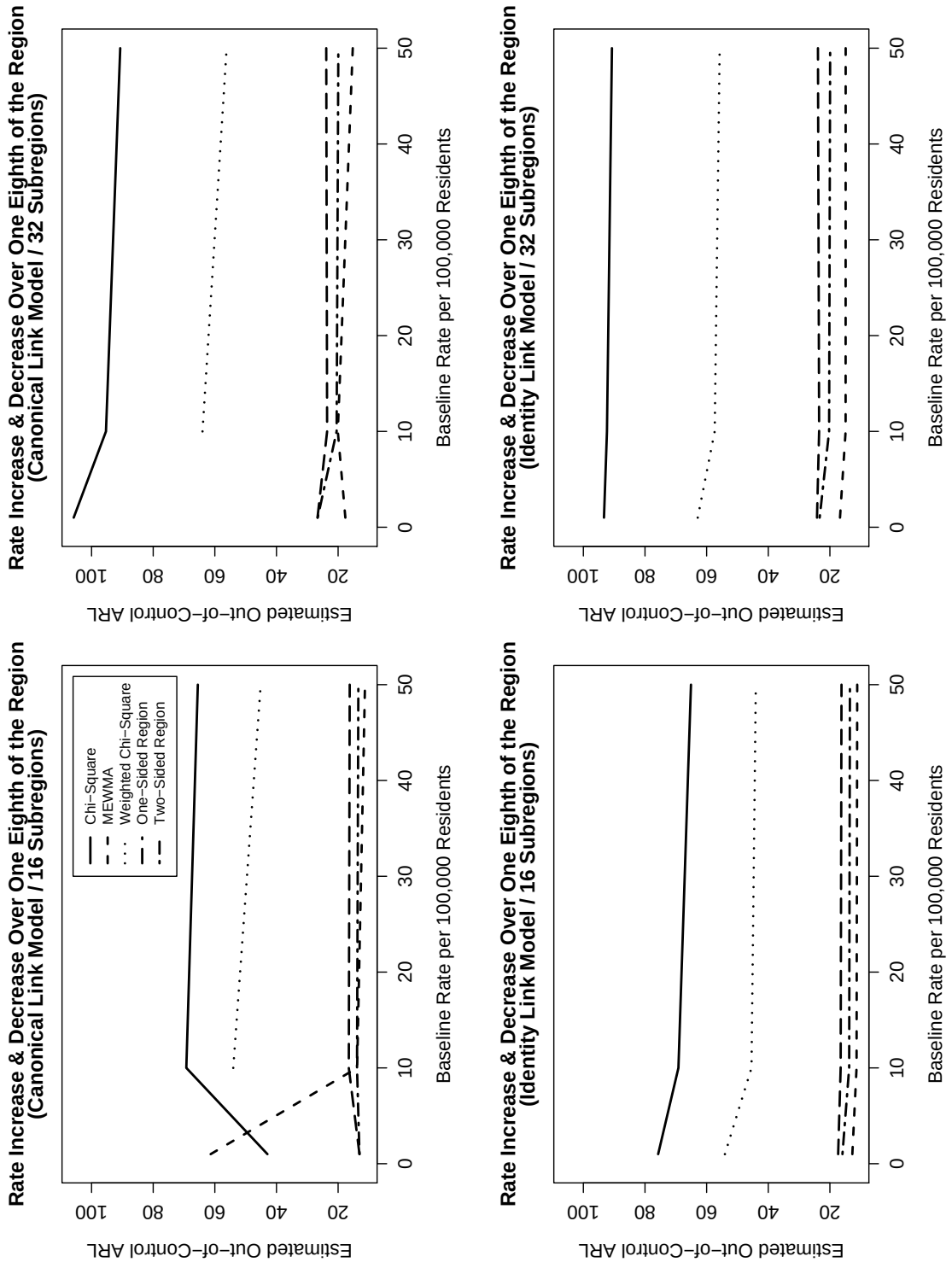
Out-of-Control ARL Results of the Wavelet-Based Disease Surveillance Method

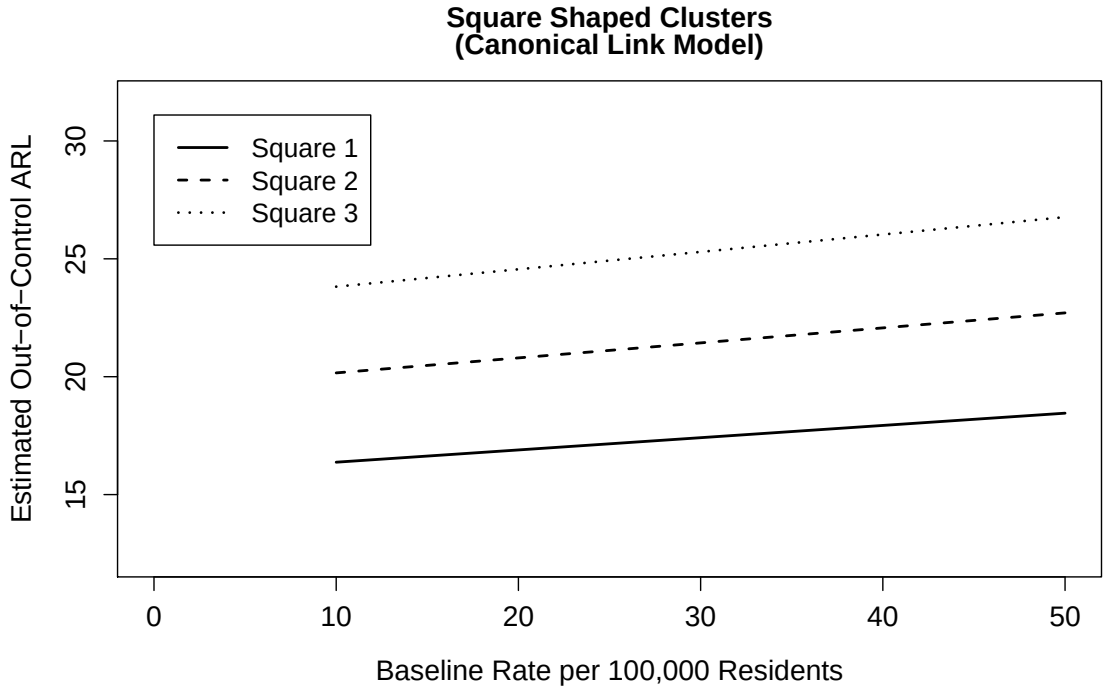
Figures of the out-of-control ARL estimates for the control charts used in the wavelet-based disease surveillance method shown in Chapter 4, Section 4.2.2 are presented in this Appendix. These figures are provided so that the patterns in out-of-control ARL performance of the control charts can be examined across different out-of-control scenarios. The first set of figures presented show the out-of-control ARL results for the out-of-control scenarios with differing region and cluster sizes. The values plotted in these figures are those given in Tables 4.6 and 4.7. The second set of figures presented show the out-of-control ARL results for the out-of-control scenarios where clusters of different shape and location were considered. The values plotted in these figures are the ARL results for the Weighted χ^2 control chart given in Tables 4.9 and 4.10.

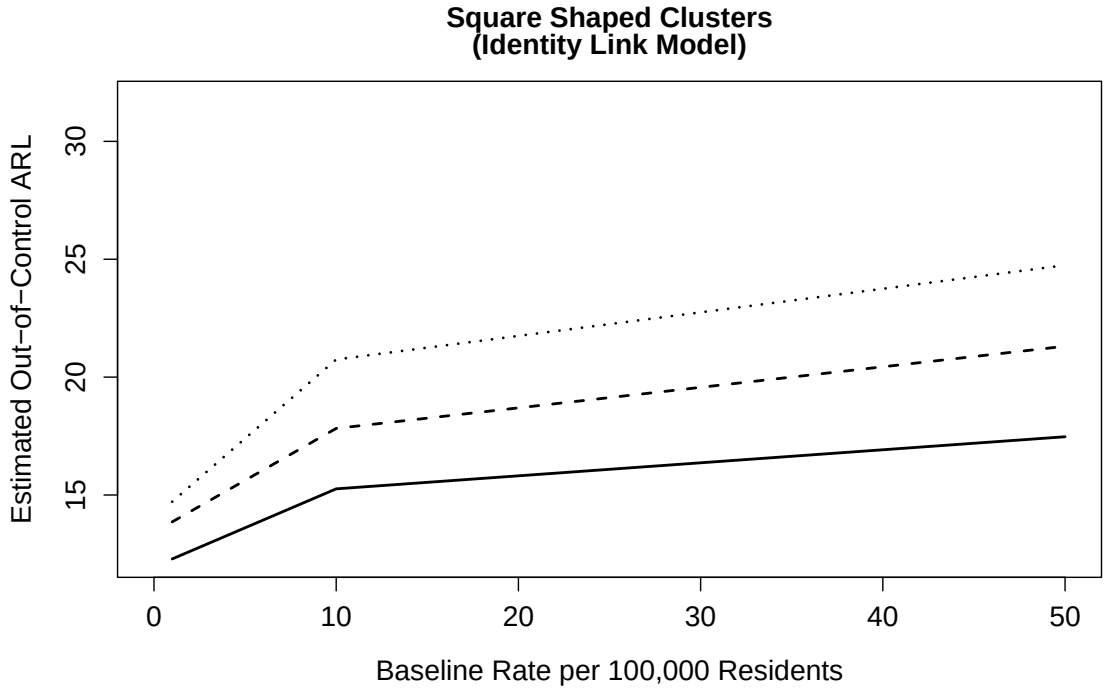

Rate Increase Over Entire Region
(Canonical Link Model / 16 Subregions)
Rate Increase Over Entire Region
(Canonical Link Model / 32 Subregions)
Rate Increase Over Entire Region
(Identity Link Model / 16 Subregions)
Rate Increase Over Entire Region
(Identity Link Model / 32 Subregions)
Chi−Square
MEWMA
Weighted Chi−Square
One−Sided Region
Two−Sided Region
Estimated Out−of−Control ARL
Baseline Rate per 100,000 Residents
0
10
20
30
40
50
50
100
150

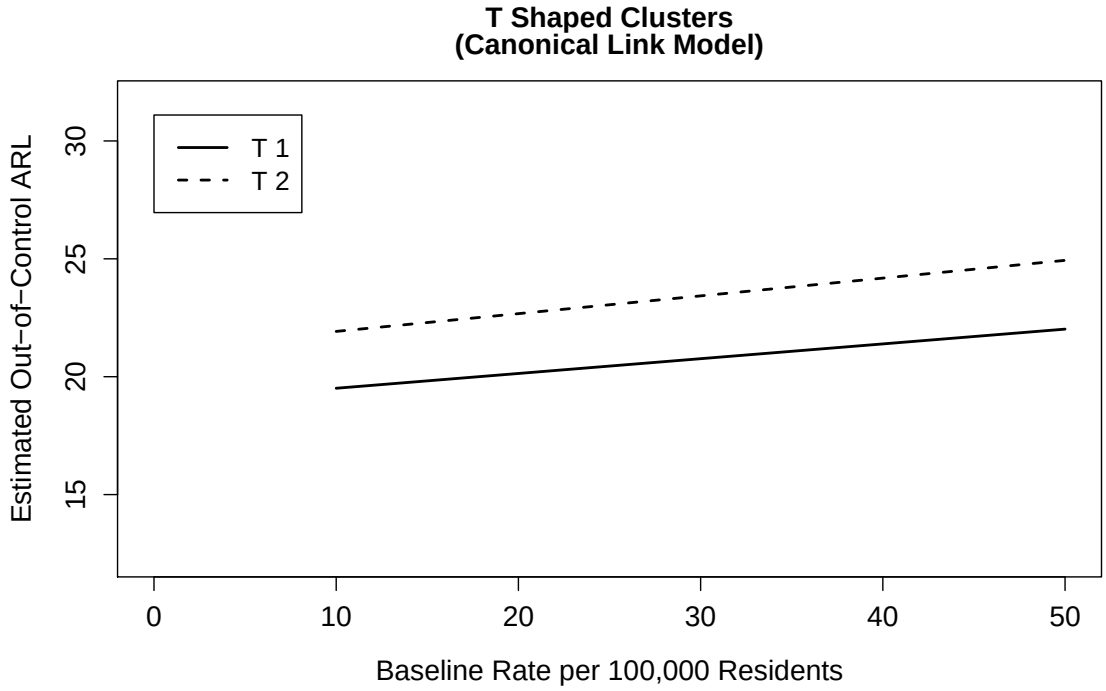

Rate Increase Over One Half of the Region
(Canonical Link Model / 16 Subregions)
Rate Increase Over One Half of the Region
(Canonical Link Model / 32 Subregions)
Rate Increase Over One Half of the Region
(Identity Link Model / 16 Subregions)
Rate Increase Over One Half of the Region
(Identity Link Model / 32 Subregions)
Chi−Square
MEWMA
Weighted Chi−Square
One−Sided Region
Two−Sided Region
Estimated Out−of−Control ARL
Baseline Rate per 100,000 Residents
10
20
30
40
50
60
70
0
10
20
30
40
50

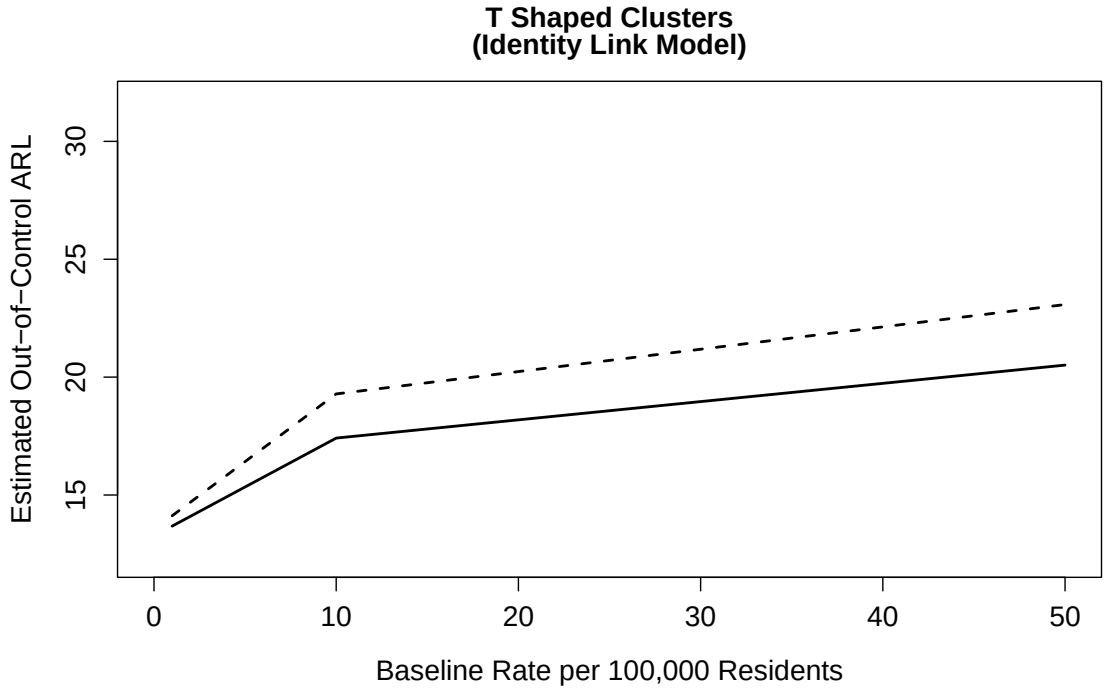

Rate Increase Over One Eighth of the Region
(Canonical Link Model / 16 Subregions)
Rate Increase Over One Eighth of the Region
(Canonical Link Model / 32 Subregions)
Rate Increase Over One Eighth of the Region
(Identity Link Model / 16 Subregions)
Rate Increase Over One Eighth of the Region
(Identity Link Model / 32 Subregions)
Chi−Square
MEWMA
Weighted Chi−Square
One−Sided Region
Two−Sided Region
Estimated Out−of−Control ARL
Baseline Rate per 100,000 Residents
10
20
30
40
50
60
70
0
10
20
30
40
50

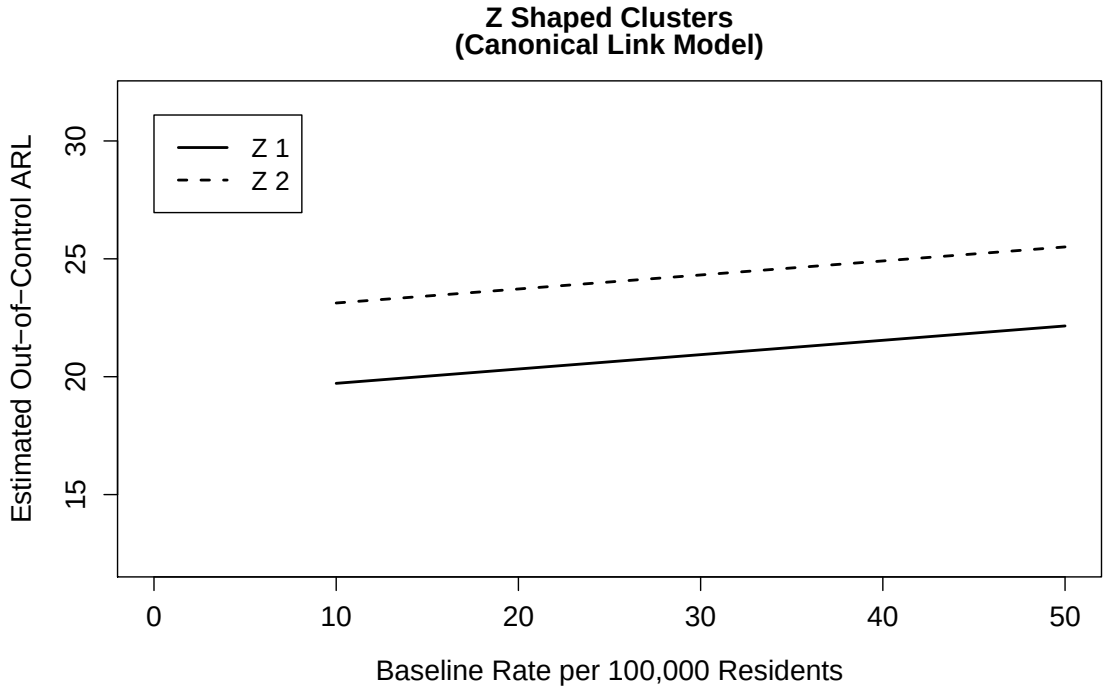

Rate Increase & Decrease Over One Eighth of the Region
(Canonical Link Model / 16 Subregions)
Rate Increase & Decrease Over One Eighth of the Region
(Canonical Link Model / 32 Subregions)
Rate Increase & Decrease Over One Eighth of the Region
(Identity Link Model / 16 Subregions)
Rate Increase & Decrease Over One Eighth of the Region
(Identity Link Model / 32 Subregions)
Chi−Square
MEWMA
Weighted Chi−Square
One−Sided Region
Two−Sided Region
Estimated Out−of−Control ARL
Baseline Rate per 100,000 Residents
20
40
60
80
100
0
10
20
30
40
50

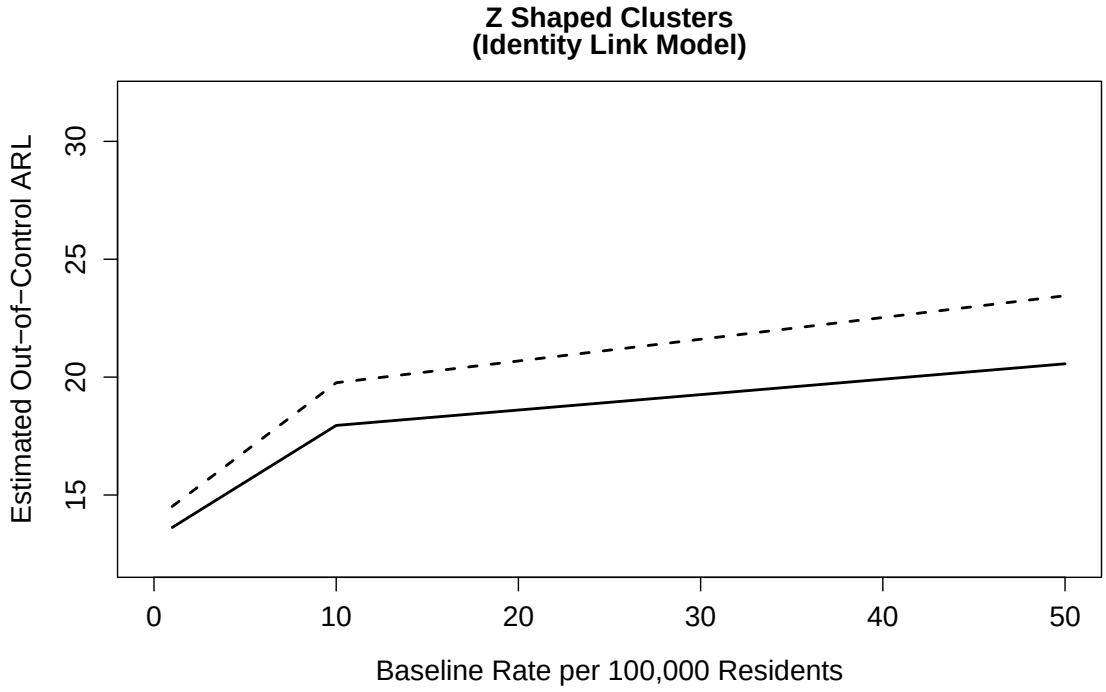

Square Shaped Clusters
(Canonical Link Model)
Square 1
Square 2
Square 3
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents

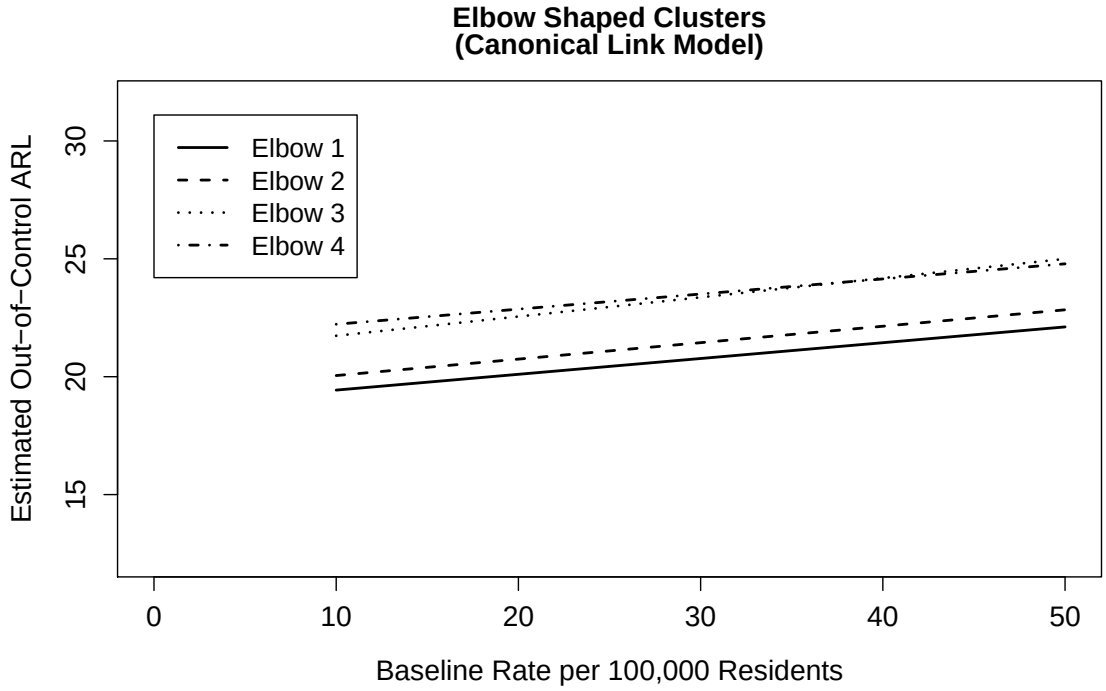

Square Shaped Clusters
(Identity Link Model)
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents

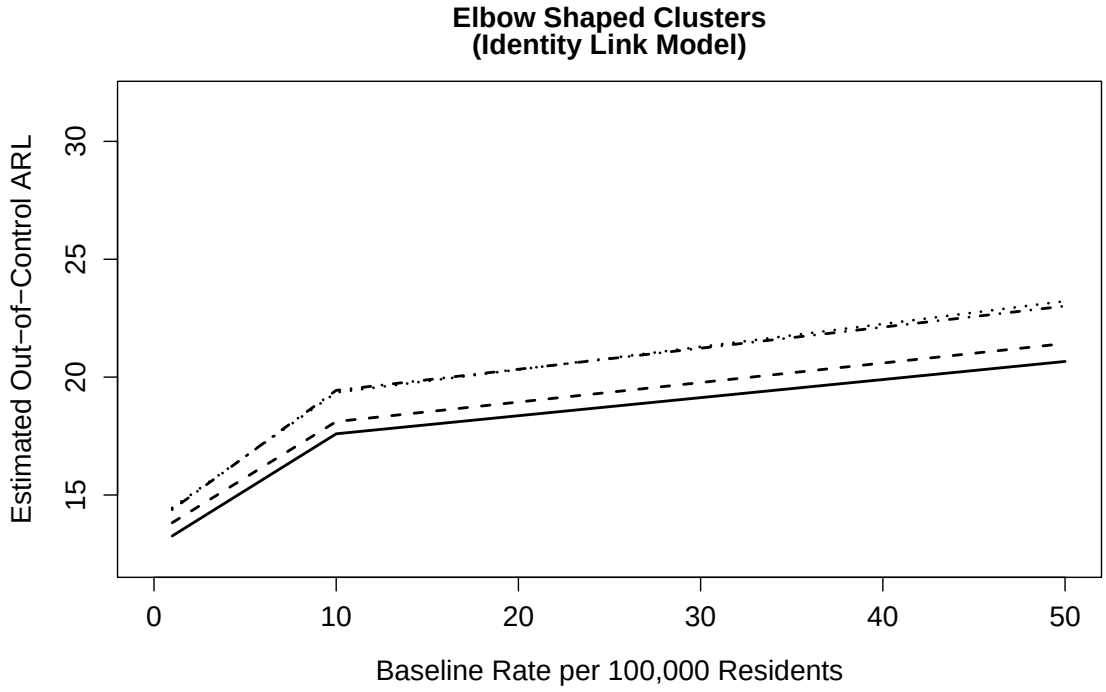

T Shaped Clusters
(Canonical Link Model)
T 1
T 2
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents

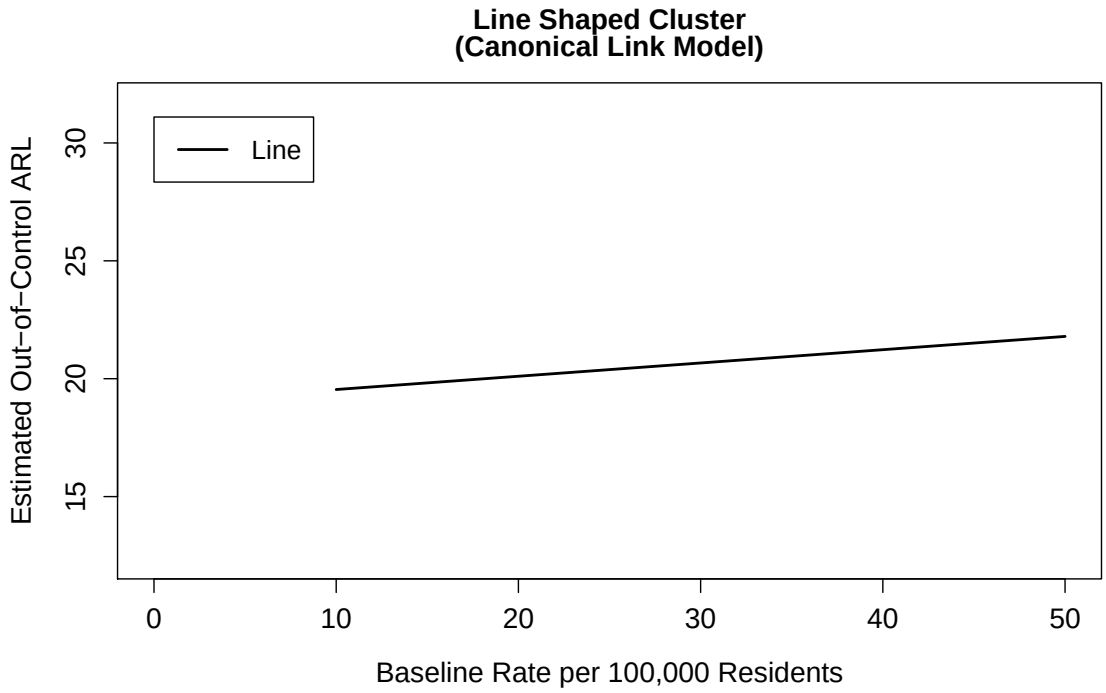

T Shaped Clusters
(Identity Link Model)
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents

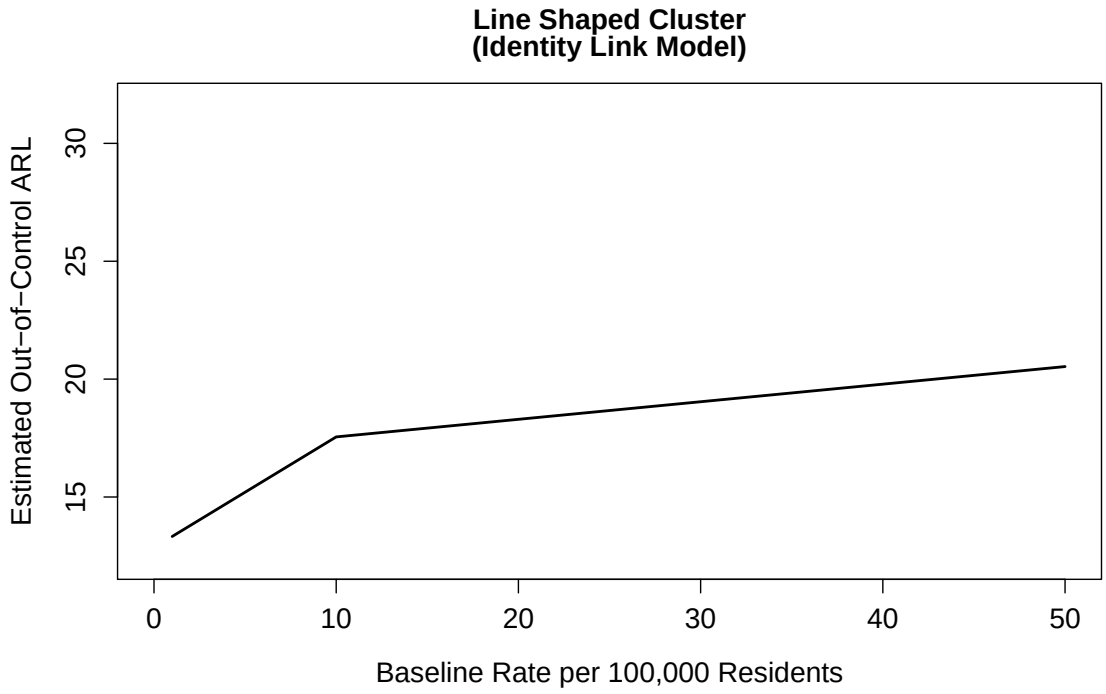

Z Shaped Clusters
(Canonical Link Model)
Z 1
Z 2
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents


Z Shaped Clusters
(Identity Link Model)
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents


Elbow Shaped Clusters
(Canonical Link Model)
Elbow 1
Elbow 2
Elbow 3
Elbow 4
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents


Elbow Shaped Clusters
(Identity Link Model)
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents


Line Shaped Cluster
(Canonical Link Model)
Line
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents


Line Shaped Cluster
(Identity Link Model)
Estimated Out−of−Control ARL
15
20
25
30
0
10
20
30
40
50
Baseline Rate per 100,000 Residents

References

Akaike, H. (1974) A new look at the statistical identification model. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Chan, T. F. and Shen, J. (2005) *Image Processing and Analysis: Variational, PDE, Wavelet, and Stochastic Methods*. Philadelphia: Society for Industrial and Applied Mathematics.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Diggle, P., Rowlingson, B. and Su, T. (2005) Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–434.

Fan, J. (1996) Tests of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, **91**, 674–688.

Farrington, C. P. and Beale, A. D. (1998) The detection of outbreaks of infectious disease. *GEOMED 1997: Proceedings of the International Workshop on Geomedical Systems*, eds. Gierl, L., Cliff, A. D., Valleron, A.-J., Farrington, P., and Bull, M. Stuttgart: Teubner, 97–117.

Goovaerts, P. and Jacquez, G. M. (2005) Moran's I and geostatistically simulated spatial neutral models. *Journal of Geographical Systems*, **7**, 137–159.

Hawkins, D. M. and Olwell, D. H. (1998) *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer.

Heymann, D. L. and Rodier, G. R. (1998) Global Surveillance of Communicable Diseases. *Emerging Infectious Diseases*, **4**, 362–365.

Holland, J. H. (1992) *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Cambridge: MIT Press.

Jacquez, G. M. (1996) A k nearest neighbour test for space-time interaction. *Statistics in Medicine*, **15**, 1935–1949.

Järpe, E. (1999) Surveillance of the interaction parameter of the Ising model. *Communications in Statistics Theory and Methods*, **28**, 3009–3027.

Jensen, D. R. and Solomon, H. (1972) A gaussian approximation to the distribution of a definite quadratic form. *Journal of the American Statistical Association*, **67**, 898–902.

Jeong, M. K., Lu, J.-C. and Wang, N. (2006) Wavelet-based SPC procedure for complicated functional data. *International Journal of Production Research*, **44**, 729–744.

Joner, Jr., M. D., Woodall, W. H., Reynolds, Jr., M. R. and Fricker, Jr., R. D. (2008) A one-sided MEWMA chart for health surveillance. *Quality and Reliability Engineering International*, **24**, 503–518.

Kleinman, K. (2005) Generalized linear models and generalized linear mixed models for small-area surveillance. *Spatial & Syndromic Surveillance for Public Health*, Chapter 5, eds. Lawson, A. B. and Kleinman, K. New York: Wiley.

Knox, E.G. (1964) The detection of space-time interactions. *Applied Statistics*, **13**, 25–29.

Kulldorff, M. (2001) Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society Series A*, **164**, 61–72.

Kulldorff, M. and Hjalmars, U. (1999) The Knox method and other tests for space-time interaction. *Biometrics*, **55**, 544–552.

Lawson, A. B. (2001) Comments on the papers by Williams *et al.*, Kulldorff, Knorr-Held and Best, and Rogerson. *Journal of the Royal Statistical Society Series A*, **164**, 97–99.

Lawson, A. B. (2006) *Statistical Methods in Spatial Epidemiology* 2nd ed. New York: Wiley.

Lawson, A. B. and Kleinman, K. (2005) *Spatial & Syndromic Surveillance for Public Health*. New York: Wiley.

Lowry, C. A., Woodall, W. H., Champ, C. W. and Rigdon, S. E. (1992) A multivariate exponentially weighted moving average control chart. *Technometrics*, **34**, 46–53.

Leung, C. S., Patel, M. S. and McGilchrist, C. A. (1999) A distribution-free regional cumulative sum for identifying hyperendemic periods of disease incidence. *The Statistician*, **48**, 215–225.

Louie, M. M. and Kolaczyk, E. D. (2006) Multiscale detection of localized anomalous structure in aggregate disease incidence data. *Statistics in Medicine*, **25**, 787–810.

Mallat, S. (1999) *A Wavelet Tour of Signal Processing* 2nd ed. San Diego: Academic Press.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Marshall, J. B., Spitzner, D. J. and Woodall, W. H. (2007) Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time. *Statistics in Medicine*, **26**, 1579–1593.

Marshall, R. J. (1991) A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society Series A*, **154**, 421–441.

Neill, D. B. (2009) An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics*, **8**:20, www.ij-healthgeographics.com.

Neill, D. B. and Cooper, G. F. (2009) A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning*, *in press*.

Ogden, R. T. (1997) *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhäuser.

Patnaik, P. B. (1949) The noncentral χ^2 and F-distributions and their applications. *Biometrika*, **36**, 202–232.

Pearson, E. S. (1959) Note on an approximation to the distribution of noncentral χ^2. *Biometrika*, **46**, 364.

Raubertas, R. F. (1989) An analysis of disease surveillance data that uses geographic locations of the reporting units. *Statistics in Medicine*, **8**, 267–271.

Reis, M. S. and Saraiva, P. M. (2006) Multiscale statistical process control of paper surface profiles. *Quality Technology and Quantitative Management*, **3**, 263–282.

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

Rogerson, P. A. (2001) Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society Series A*, **164**, 87–96.

Rogerson, P. A. and Yamada, I. (2004) Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statistics in Medicine*, **23**, 2195–2214.

Salomon, D. (2000) *Data Compression: The Complete Reference* 2nd ed. New York: Springer-Verlag.

Shmueli, G. (2005) Wavelet-based monitoring in modern biosurveillance. *Technical Report, RHS-06-002, University of Maryland, Robert H. Smith School of Business.*

Shmueli, G. and Burkom, H. (2009) Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, *in press.*

Siegmund, D. O. (1985) *Sequential Analysis: Tests and Confidence Intervals.* New York: Springer.

Sonesson, C. (2007) A CUSUM framework for detection of space-time disease clusters using scan statistics. *Statistics in Medicine*, **26**, 4770–4789.

Sonesson, C. and Bock, D. (2003) A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A*, **166**, 5–21.

Spitzner, D. J. (2008) A powerful test based on tapering for use in functional data analysis. *Electronic Journal of Statistics*, **2**, 939–962.

Spitzner, D. J. and Marshall, J. B. (2008) Directed spatio-temporal monitoring of disease incidence surfaces. *Presented at the Spring Meetings of the Eastern North American Region of the International Biometric Society*, Arlington, VA, March 17, 2008.

Tango, T. and Takahashi, K. (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**:11, www.ij-healthgeographics.com.

Thacker, S. B., Stroup, D. F., Rothenberg, R. B. and Brownson, R. C. (1995) Public health surveillance for chronic conditions: a scientific basis for decision. *Statistics in Medicine*, **14**, 629–641.

Vance, L. C. (1986) Average run lengths of cumulative sum control charts for controlling normal means. *Journal of Quality Technology*, **18**, 189–193.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. New York: Wiley.

Williams, E. H., Smith, P. G., Day, N. E., Geser, A., Ellice, J. and Tukei, P. (1978) Space-time clustering of Burkitt's lymphoma in the West Nile District of Uganda: 1961–1975. *British Journal of Cancer*, **37**, 109–122.

Williams, G. W. (1984) Time-space clustering of disease. *Statistical Methods for Cancer Studies*, Chapter 5, ed. Cornell, R. G. New York: Marcel Dekker.

Wilson, E. B. and Hilferty, M. M. (1931) The distribution of chi-square. *Proceedings of the National Academy of Sciences*, **17**, 684–688.

Woodall, W. H. (2006) The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, **38**, 89–104.

Woodall, W. H., Marshall, J. B., Joner, Jr., M. D., Fraker, S. E., and Abdel-Salam, A.-S. G. (2008) On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of the Royal Statistical Society Series A*, **171**, 223–237.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C. and Gupta, S. (2004) Using control charts to monitor process and product quality profiles. *Journal of Quality Technology*, **36**, 309–320.

Zhou, H. and Lawson, A. B. (2008) EWMA smoothing and Bayesian spatial modeling for health surveillance. *Statistics in Medicine*, **27**, 5907–5928.

Zhou, S., Baocheng, S. and Shi, J. (2006) An SPC monitoring system for cycle-based waveform signals using Haar transform. *IEEE Transactions on Automation Science and Engineering*, **3**, 60–72.