An Examination of Spatial Scan Statistics for Time to Event Data

by

Iram Usman

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

The spatial scan statistic (SSS) has been used for the identification of geographical clusters of higher than expected numbers of cases of a condition such as an illness. Disease outbreaks in a geographic area are a typical example. These statistics can also identify geographic areas with longer time to events if the SSS uses appropriate distribution. Other authors have proposed the exponential and Weibull distributions for the event times. We have established the log-Weibull distribution as a new and alternative approach for the SSS, and compared and contrasted the three distributions through simulation studies to investigate right censoring. Different datasets from the exponential, Weibull, log-Normal, and gamma probability distributions have been generated in order to test the robustness of the SSS's. Three differential censoring settings were imposed on the generated datasets to test the detection power of the true spatial cluster by each SSS. The method along with the existing exponential and Weibull SSS's were also illustrated on the time to specialist visit (cardiology or internal medicine) data for discharged patients presenting to an Emergency Department for atrial fibrillation and flutter in Alberta during 2010-2011.

Disclaimer

This study is based in part on data provided by Alberta Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the Government nor Alberta Health express any opinion in relation to this study.

Acknowledgments

First of all, I would like to express my thanks to my supervisor, Dr. Rhonda Rosychuk for her great and gracious support throughout the thesis. She has always been very efficient and kind to me, and her extensive professional knowledge and experience in the field of Biostatistics provided me countless help to finish this task. In addition, I would like to thank Dr. Rohana J. Karunamuni for his support and guidance. He provided me with invaluable advice during my two years of study at the University of Alberta.

I pay my thanks to Alberta Health for extracting the data. This study is funded by Dr. Rhonda Rosychuk from her NSERC grant. Dr. Rosychuk is salary supported by Alberta Innovates-Health Solutions as a Health Scholar.

I gratefully acknowledge all of the unconditional support from my family and friends, who have always been very kind and supportive to me. Last but not least, I cannot express enough thanks to my husband, Usman Zafar, for being very supportive and helpful throughout my studies. Without his help, I might not have been able to finish this task. I dedicate this work to him with love and special thanks.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction and Background

A cluster is defined as a group of objects or people with some common characteristics. The existence of more than a presumed number of cases of a disease condition, such as a disease outbreak within a certain region, is referred to as a spatial disease cluster.[1] Public health officials, epidemiologists, and researchers use various cluster detection techniques to test for the presence and location of the occurrence of incident cases of diseases, that comprise spatial disease clusters. These officials seek to find the geographical clusters which are least likely to appear by chance with an elevated number of disease cases. Early and timely detection of spatial disease clusters enables the health authorities to take actions to assist in understanding the distribution of disease and if possible, control disease.

A large number of methods have been proposed and applied by authors for the identification and evaluation of geographical disease clusters and for disease surveillance. These methods include the scan statistic,[2,3] global clustering tests,[4] kriging,[5,6] and spatial smoothing methods.[7] Kulldorff and Nagarwalla proposed the scan statistic in a purely spatial setting for geographic disease surveillance and the evaluation of disease cluster alarms.[8] The method was capable of identifying spatial clusters of variable sizes and locations in the study area. They applied their proposed method to leukaemia patients in Upstate New York during 1978-1982. Besag and Newell developed a spatial cluster detection method to find proximities of unexpected occurrence of certain events with a fixed number of cases.[9] They illustrated the methodology on data for acute lymphoblastic leukaemia patients in England during 1975-1985.

Petrisor et al. used the kriging and the DAC (difference between the empirical cumulative distribution of cases and non-cases at a particular point) statistic to predict the low birthweight clusters in Spartanburg county, SC, United States during 1989-1990.[10] Thomas and Carlin used the technique of data smoothing using Bayesian methods implemented via Markov chain Monte Carlo, in addition to the various mapping and cluster detection techniques on the breast and colorectal cancer detection data for Minnesota counties for 1995-1997.[7]

This study focuses on the scan statistic in two dimensions, i.e., the spatial scan statistic (SSS). The SSS has been widely used as a standardized approach with its possible extensions for the last two decades, not only in the disease clustering literature but also in various other fields of study like natural disasters,[11] forestry,[12] astronomical data,[13] history,[14] and psychology.[15] The key reasons for the popularity of this method include that it identifies the cluster location and tests the tendency to cluster.[16] According to Costa and Assunção, the latter advantage is considered to be more important in terms of health related interventions than global clustering results.[16] Another important reason for the popularity of the SSS is that it can be easily analyzed in SaTScan, which is an easy to operate free software, originally developed by Martin Kulldorff.[17] SaTScan allows the choice of using different probability models for the SSS including the Poisson, Bernoulli, Normal, and the exponential.

The SSS's based on the Bernoulli and Poisson models are frequently used for count data for cluster identification and geographical disease surveillance.[18] These scan statistics have been further extended to other kinds of data such as ordinal,[20] multinomial,[21] and continuous data.[22]

The SSS for time to event data is used to determine if there are geographical clusters with either longer than expected and/or shorter than expected times to event. For example, in case of a survival time, if the SSS can detect the geographical clusters of people with shorter than expected survival time, it may be an indication of antagonistic or insufficient treatment or health practices.[19] Similarly, if longer than expected survival time geographical clusters are detected, it may reflect advancement of the treatment or better health conditions of the people.[19] The SSS's based on the exponential[18] and Weibull[19] probability models have already been proposed by other authors. We propose the log-Weibull as an alternative distribution for the SSS for cluster detection of time to event data. The log-Weibull distribution has wide applications in extreme value theory. Our focus is to establish a new SSS for the cluster detection of rare and extreme events.

1.2 Objectives of the Study

Main objectives of this study are to:

1. Develop a spatial scan statistic based on the log-Weibull distribution for detecting spatial clusters.

2. Use the new method to identify clusters of longer times to specialist physician follow-up after an emergency department presentation for atrial fibrillation and flutter in Alberta, Canada.

3. Compare the new method with methods based on the exponential[18] and Weibull[19] spatial scan statistics for the Alberta data.

4. Perform simulation studies to investigate power, the effect of right (type I) differential censoring, and the ability to identify the true cluster.

1.3 Structure of the Chapters

Brief methodologies for the general SSS and SSS's based on the exponential, Weibull and log-Weibull distributions are described in Chapter 2. The new developed method along with the existing exponential and Weibull SSS's have been illustrated on the administrative data from Alberta Health and the results are presented in Chapter 3. Chapter 4 contains the results from the simulation studies performed to compare and contrast the three SSS's to investigate the effect of right differential censoring on the power of detection of a potential cluster. A discussion and suggested future work follow in Chapter 5.

Chapter 2

Methodology

2.1 Overview

A brief review of the literature on the spatial scan statistic (SSS) is first provided. Next, the details of a new formulation for the SSS using the log-Weibull distribution are given after the descriptions of the SSS's based on the exponential and Weibull distributions. The chapter ends with a discussion of a permutation test to obtain p-values and a summary.

2.2 Literature Review

A tremendous amount of work has been done on the SSS since it was first presented by Naus[3], in his paper entitled "clustering of random points in two dimensions". In this section, some of the literature based on the SSS and its extensions has been reviewed. The main focus of the review is on the application, various data types and models, comparison with other methods, and the advancement of the SSS.

Kulldorff and Nagarwalla developed the SSS for the detection and evaluation of potential spatial clusters for a specified event such as a disease outbreak.[8] The proposed technique was based on the Bernoulli model. Since then, many advances have handled different data types and statistical models.

Kulldorff presented the general statistical theory of the SSS for count data based on the Bernoulli and Poisson distributions in his published research paper "A spatial scan statistic".[24] He illustrated the method on sudden infant death syndromes in North Carolina. The theory of the SSS was further extended by Kulldorff et al. for continuous data by using the Normal distribution.[22] The authors used the proposed methodology to detect the geographical clusters of low birth weight and early gestation in New York city.

The SSS has also been constructed for ordinal,[20] multinomial,[21] and multivariate[25] data types. The SSS's based on these three data types were applied on the prostate cancer grade and stage data in the Maryland,[20] meningitis data in the United Kingdom,[21] and data from the National Bioterrorism Syndromic Surveillance Demonstration Project in Massachusetts,[25] respectively. The Weighted Normal SSS has also been developed based on the heterogeneous population data and was applied to the short-term and long-term lung cancer survival data in Los Angeles County.[26] Rosychuk and Chang developed the SSS with a compound Poisson model for the correlated count data.[27] They illustrated the proposed method on multiple disease-related visits to emergency departments in Alberta.

Time to event data (e.g., survival time) is one of the important health outcomes for which the SSS is of deep interest.[18] Censoring is an important and main characteristic of time to event data, which may occur when the data is partially or not available due to some reason. The SSS has the capability of incorporating the censored data. The exponential[18] and Weibull[19] SSS's (adjusted for censoring) have been developed for time to event data. These SSS were illustrated on survival data for men diagnosed with prostate cancer in Connecticut[18] and tuberculosis patients in Nainital district of Uttarakhand, India,[19] respectively.

The SSS has been further advanced to three dimensions by incorporating time as a third dimension, and is called the space-time scan statistic. Kulldorff et al. introduced the retrospective space-time scan statistic and illustrated the method on the evaluation of a brain cancer cluster alarm in Los Alamos, NM.[28] Kulldorff also proposed a prospective space-time scan statistic and applied the method on the data of thyroid cancer among men in New Mexico.[29]

Many other authors have developed spatial and space-time scan statistics for non-circular, irregular, and flexible scanning windows. Research on these windows include that of Kulldorff et al., who introduced the SSS for an elliptical scanning window of variable location, shape, angle, and size.[30] They used this technique on the breast cancer mortality data from Northeastern United States and female oral cancer mortality in the United States. Iyenger extended the restrictive shaped space time clustering technique to a flexible square pyramid base and applied the method on a real dataset of brain cancer occurrence over a 19 year period.[31] Takashi et al. presented a flexible space-time scan statistic and showed that it is more appropriate for detecting and monitoring disease outbreaks in irregularly shaped areas.[32] They illustrated the proposed method on daily syndromic surveillance data in eastern Massachusetts.

Many authors have performed the comparison among different cluster detection methodologies. For example, Torabi and Rosychuk compared five popular spatial disease clustering methods by analyzing the dataset of malignant cancer diagnoses in children in Alberta during 1983-2004.[33] The five methods under consideration were the Besag-Newell, the circular spatial scan statistic, the flexible spatial scan statistic, the Tango's maximized excess events test, and the Bayesian disease mapping. The authors concluded that the performance of all these methods was good, but the Besag-Newell and Tango's maximized excess events test methods were able to identify both local (region of interest) and general (any significant cluster in the study region) clusters. Costa and Assunção performed a fair comparison between two of the most popular clustering methods, i.e., the spatial scan and the Besag-Newell disease clustering tests.[16] A publicly available simulated benchmark dataset was used to find that both of the methods produced the similar results, except the performance of the spatial scan was better for the clusters located in regions with sparse population.

Although, spatial and space-time scan statistics have gained wide popularity and have been applied successfully in numerous fields of study, researchers are developing new techniques to improve the accuracy and computational efficiency of spatial cluster detection. Neil have proposed a fast subset scan approach based on the "linear time subset scanning property", and proved that this approach significantly improves the timeliness and accuracy of the event detection.[34] It was demonstrated that the proposed technique detected the disease outbreaks two days faster than the other detection methods.

2.3 The Methodology of the Spatial Scan Statistic

The SSS is a statistical technique for identifying the geographic zones from a study region that have the strongest indication of representing a spatial cluster. The SSS can be used for both, spatially aggregated data and data for the exact geographic co-ordinates for each individual.[8] For aggregated data, the study region is divided into non-overlapping geographical sub-regions, each characterized by a centroid (either geographical or population based) and the data are aggregated to the sub-region's centroid. In case of non-aggregated data, each region contains only one individual.[8]

The SSS for cluster detection uses data such as administrative health data collected for geographical sub-regions. The SSS imposes circular searching window (also called circular spatial scan window) of radius r on each centroid with its center at the co-ordinate of a centroid.[8] A zone (Z) defined by this circular window is comprised of all the individuals in those sub-regions whose centroids lie inside the circle.[8] For the purpose of the analysis, an upper bound r^* (usually between 10% and 50% of the total population) is chosen for the radius of the circular spatial scan window.[19]

For each region's centroid, its nearest neighbours covering altogether r^* percent of the total population are calculated. For any given position of the centroid, the radius of the window is expanded continuously to take any value between

0 and r^*.[19] During the expansion, every time a new zone is created with an inclusion of a new neighbouring centroid in the circular window.[27] In this whole process, there would be infinite number of distinct circles, but a finite number of zones.[8] Zones defined in this way have irregular geographical boundaries depending on the size and shape of those sub-regions, whose centroids lie inside the spatial scan window.[27]

The methodology of SSS is based on calculating the maximum likelihood ratio, or more precisely maximum log likelihood ratio (LLR). The SSS partitions the geographical area into zones (i.e., areas of potential cluster versus the rest of the study region) and the LLR is calculated every time when a new zone is created for each centroid.[19,24] The zone maximizing the LLR is called the most likely cluster (i.e., the cluster least likely to occur by chance).

Let the most likely cluster be the zone $\hat{Z}$ that maximizes the LLR. The hypothesis under consideration is:

- $\triangleright$ H_0: The disease risk is constant over $\hat{Z} \cup \hat{Z}^c$.
- $\triangleright$ H_1: There is an elevated risk in $\hat{Z}$.

Since information about the exact statistical distribution of the test statistic is not known, a permutation testing procedure is used to perform the hypothesis testing. The associated p-value is calculated to check the statistical significance of the potential cluster. The following subsections describe the methodologies and test statistics for the already developed SSS's based on the exponential[18] and Weibull[19] distributions along with the newly developed SSS based on the log-Weibull distribution. For symmetry, the same notations (Section 2.3.1) have been used to describe the three SSS's.

2.3.1 Notations

Let G be the whole study region which can be partitioned into Z and Z^c mutually exclusive sub-regions, where Z indicates a zone designated to be a potential cluster and Z^c is the rest of the study region. The null hypothesis of existence of no cluster for any Z is contrasted with one of three alternative hypotheses:

- ▷ At least one zone is detected with shorter than expected times to event.
- ▷ At least one zone is detected with longer than expected times to event.
- ▷ At least one zone is detected with either shorter or longer than expected times to event.

Let $N = n_{in} + n_{out}$ be the total number of individuals in G, where n_{in} and n_{out} are the total individuals inside and outside the zone, respectively. The subscript "*in*" indicates the object is calculated from the individuals inside the zone. Similarly, the subscript "*out*" indicates the measurement is calculated from the individuals outside the zone.

Let the i^{th} individual have a time to event T_i, $(i = 1, ..., N)$ or a fixed right censoring time L_i. The event time T_i is observed if $T_i \leq L_i$ $(\delta_i = 1)$, and L_i is observed if $T_i > L_i$ $(\delta_i = 0)$, where δ_i is the indicator to represent if time is censored or not.[18] The observed time is defined as $t_i = \min(T_i, L_i)$. Let $R = r_{in} + r_{out}$ be the total number of uncensored observations, where r_{in} and

r_{out} are the total number of uncensored observations inside and outside the zones, respectively. These are defined as:

$$r_{in} = \sum_{i \epsilon Z} \delta_i \qquad \text{and} \qquad r_{out} = \sum_{i \epsilon Z^c} \delta_i \ .$$

2.3.2 A Spatial Scan Statistic Based on the Exponential Distribution

Huang et al. have proposed the SSS based on the exponential model for continuous time to event (survival) data, which has the ability to incorporate both censored and uncensored observations.[18] In this section, a brief review of the SSS based on the exponential distribution has been presented. Complete details of the methodology can be found in the paper presented by Huang et al.[18]

Assume that the times to event T_i's $(i = 1, ..., N)$ are independently and identically distributed (i.i.d.) with the exponential probability density function (PDF)

$$f(T_i) = \frac{1}{\theta}\, e^{-T_i/\theta} \qquad\qquad T_i \geq 0, \tag{2.3.1}$$

where θ is the scale parameter. It is further assumed that the time to event for each individual inside the zone is distributed as the exponential distribution with the scale parameter θ_{in}. Similarly, the times to event for individuals outside the zone are also exponentially distributed with the scale parameter θ_{out}. The mathematical forms of the null and alternative hypotheses of shorter, longer, and either shorter or longer than expected times to event are

$H_0 : \theta_{in} = \theta_{out}$, $H_1 : \theta_{in} < \theta_{out}$, $H_1 : \theta_{in} > \theta_{out}$, and $H_1 : \theta_{in} \neq \theta_{out}$, respectively.

The likelihood function, based on the exponential distribution for any zone Z can be written as

$$L(Z, \theta_{in}, \theta_{out}) = \frac{1}{(\theta_{in})^{r_{in}}}\, e^{-\sum_{i \epsilon Z}\left(\frac{t_i}{\theta_{in}}\right)} \frac{1}{(\theta_{out})^{r_{out}}}\, e^{-\sum_{i \epsilon Z^c}\left(\frac{t_i}{\theta_{out}}\right)}. \tag{2.3.2}$$

The likelihood ratio test statistic for the the alternative hypothesis $H_1 : \theta_{in} \neq \theta_{out}$ for at least one zone Z is

$$\lambda = \frac{\max_{Z, \theta_{in} \neq \theta_{out}} L\left(Z, \theta_{in}, \theta_{out}\right)}{\max_{Z, \theta_{in} = \theta_{out}} L\left(Z, \theta_{in}, \theta_{out}\right)} = \frac{L(\hat{Z})}{\hat{L}} \tag{2.3.3}$$

where $\hat{Z}$ is the zone maximizing $L(Z, \theta_{in}, \theta_{out})$ under H_1, and $\hat{L}$ is the maximum of $L(Z, \theta_{in}, \theta_{out})$ under H_0. After applying the natural log on $L(Z, \theta_{in}, \theta_{out})$ and taking the derivatives, the maximum likelihood estimates of θ_{in} and θ_{out} are $\hat{\theta}_{in} = \dfrac{r_{in}}{\sum_{i \epsilon Z} t_i}$ and $\hat{\theta}_{out} = \dfrac{r_{out}}{\sum_{i \epsilon Z^c} t_i}$, respectively.

By using the maximum likelihood estimates of the scale parameters for inside and outside the zone, the derived likelihood ratio test statistic for the exponential SSS for $H_1 : \theta_{in} \neq \theta_{out}$ is

$$\lambda = \frac{\max_Z \left(\dfrac{r_{in}}{\sum_{i \epsilon Z} t_i}\right)^{r_{in}} \left(\dfrac{r_{out}}{\sum_{i \epsilon Z^c} t_i}\right)^{r_{out}}}{\left(\dfrac{R}{\sum_{i \epsilon G} t_i}\right)^{R}}. \tag{2.3.4}$$

The test statistic λ in (2.3.4) is multiplied by $I\left(\dfrac{r_{in}}{\sum_{i \epsilon Z} t_i} < \dfrac{r_{out}}{\sum_{i \epsilon Z^c} t_i}\right)$, if at least one cluster is to be detected with shorter than expected times to event.

Similarly, λ is multiplied by $I\left(\frac{r_{in}}{\sum_{i\epsilon Z} t_i} > \frac{r_{out}}{\sum_{i\epsilon Z^c} t_i}\right)$, if one wants to detect at least one cluster with longer than expected times to event.

2.3.3 A Spatial Scan Statistic Based on the Weibull Distribution

Bhatt and Tiwari established the SSS based on the Weibull distribution in the same spirit of the SSS based on the exponential distribution. The Weibull model is a nice generalization of the exponential model by including a shape parameter with the existing scale parameter.[19] The additional parameter provides the opportunity for the hazard function of the Weibull distribution to take different shapes rather than to be a constant like the exponential distribution. In this section, the methodology of the SSS based on the Weibull distribution has been summarized. Complete details of the methodology can be found in the paper presented by Bhatt and Tiwari.[19]

Let the times to event $T_i^{\prime}$s $(i = 1, ..., N)$ be i.i.d. with the Weibull PDF

$$f(T_i) = \frac{1}{\theta}\, p\, T_i^{(p-1)} e^{\left(-T_i^p/\theta\right)} \qquad T_i \geq 0, \tag{2.3.5}$$

where $\theta =$ scale and $p =$ shape parameters, respectively. Let the time to event for each individual inside the zone be distributed as the Weibull distribution with θ_{in} and p_{in} as the scale and shape parameters, respectively. Similarly, assume that the times to event for individuals outside the zone are Weibull distributed with θ_{out} and p_{out} as the scale and shape parameters, respectively.

The null hypothesis under consideration is $H_0 : \theta_{in} = \theta_{out}$ versus the alternative hypotheses $H_1 : \theta_{in} < \theta_{out}$, $H_1 : \theta_{in} > \theta_{out}$, or $H_1 : \theta_{in} \neq \theta_{out}$, representing the existence of spatial clusters of individuals with shorter than expected, longer than expected, or either shorter or longer than expected times to event, respectively.

By using the same statistic shown in (2.3.3), the derived likelihood ratio test statistic for the Weibull SSS for $H_1 : \theta_{in} \neq \theta_{out}$ is

$$\lambda = \max_Z \frac{\left(\frac{R}{\sum_{i\epsilon G} t_i^p}\right)^R}{\left(\frac{r_{in}}{\sum_{i\epsilon Z} t_i^{p_{in}}}\right)^{r_{in}} \left(\frac{r_{out}}{\sum_{i\epsilon Z^c} t_i^{p_{out}}}\right)^{r_{out}}} \ . \tag{2.3.6}$$

For $H_1 : \theta_{in} < \theta_{out}$, λ is multiplied by $I\left(\frac{r_{in}}{\sum_{i\epsilon Z} t_i^p} < \frac{r_{out}}{\sum_{i\epsilon Z^c} t_i^p}\right)$, and similarly for $H_1 : \theta_{in} > \theta_{out}$, it is multiplied by $I\left(\frac{r_{in}}{\sum_{i\epsilon Z} t_i^p} > \frac{r_{out}}{\sum_{i\epsilon Z^c} t_i^p}\right)$.

2.3.4 A Spatial Scan Statistic Based on the Log-Weibull Distribution

The SSS's for the exponential[18] and Weibull[19] distributions are very useful in the field of cluster detection for time to event data. These two methods help motivated the new methodology of the SSS based on the log-Weibull distribution, presented in this section. The log-Weibull distribution, also known as the Gumbel distribution, is a specialized case of the generalized extreme value distribution. It is often used to model the distribution of extreme values, strength, event history data such as quick wear-out after reaching a certain age, and logarithms of times.[36] The log-Weibull distribution has a direct relationship with the Weibull distribution by a logarithmic transformation of the Weibull random variable.[37]

Assume that times to event T_i's $(i = 1, ..., N)$ are i.i.d. with the log-Weibull PDF

$$f(T_i) = \frac{1}{b}\exp\left(\frac{T_i - a}{b}\right)\exp\left\{-\exp\left(\frac{T_i - a}{b}\right)\right\} \qquad T_i \geq 0, \tag{2.3.7}$$

where b = scale and a = location parameters. In general, the log-Weibull random variable ranges from $-\infty$ to ∞, but as the time to event can not take a negative value, its range is considered from 0 to ∞ in this case. The log-Weibull PDF has no shape parameter (i.e., its PDF has only a constant shape).[36]

The survival function for the log-Weibull distribution is

$$S(T_i) = \exp\left\{-\exp\left(\frac{T_i - a}{b}\right)\right\}. \tag{2.3.8}$$

Let the time to event for each individual inside zone Z be log-Weibull distributed with a_{in} and b_{in} as the location and scale parameters, respectively. Similarly, the time to event for each individual outside zone Z (i.e., inside Z^c) follows the log-Weibull distribution with a_{out} and b_{out} as the location and scale parameters, respectively.

The null hypothesis $H_0 : b_{in} = b_{out}$ for any Z is contrasted with one of three alternative hypotheses: $H_1 : b_{in} < b_{out}$, $H_1 : b_{in} > b_{out}$, or $H_1 : b_{in} \neq b_{out}$. The alternative hypotheses show that at least one zone is detected with either shorter than expected, longer than expected, or either longer or shorter than expected times to event. The likelihood function $L(Z) = L(Z, b_{in}, b_{out})$ for the log-Weibull SSS can be written as:

$$
\begin{aligned}
L(Z) &= \prod_{i \epsilon Z} \left[\left(f(T_i)\right)^{\delta_i} \left(S(L_i)\right)^{1-\delta_i} \right] \prod_{i \epsilon Z^c} \left[\left(f(T_i)\right)^{\delta_i} \left(S(L_i)\right)^{1-\delta_i} \right] \\
&= \prod_{i \epsilon Z} \left[\left(\frac{1}{b_{in}} e^{\left(\frac{T_i - a_{in}}{b_{in}}\right) - e^{\left(\frac{T_i - a_{in}}{b_{in}}\right)}} \right)^{\delta_i} \left(e^{-e^{\left(\frac{L_i - a_{in}}{b_{in}}\right)}} \right)^{1-\delta_i} \right] \\
&\quad \times \prod_{i \epsilon Z^c} \left[\left(\frac{1}{b_{out}} e^{\left(\frac{T_i - a_{out}}{b_{out}}\right) - e^{\left(\frac{T_i - a_{out}}{b_{out}}\right)}} \right)^{\delta_i} \left(e^{-e^{\left(\frac{L_i - a_{out}}{b_{out}}\right)}} \right)^{1-\delta_i} \right] \\
&= \prod_{i \epsilon Z} \left[\frac{1}{b_{in}^{\delta_i}} e^{\delta_i \left(\left(\frac{T_i - a_{in}}{b_{in}}\right) - e^{\left(\frac{T_i - a_{in}}{b_{in}}\right)} \right)} e^{-(1-\delta_i) e^{\left(\frac{L_i - a_{in}}{b_{in}}\right)}} \right] \\
&\quad \times \prod_{i \epsilon Z^c} \left[\frac{1}{b_{out}^{\delta_i}} e^{\delta_i \left(\left(\frac{T_i - a_{out}}{b_{out}}\right) - e^{\left(\frac{T_i - a_{out}}{b_{out}}\right)} \right)} e^{-(1-\delta_i) e^{\left(\frac{L_i - a_{out}}{b_{out}}\right)}} \right] \\
&= \left(b_{in}\right)^{-r_{in}} \left(b_{out}\right)^{-r_{out}} \\
&\quad \times e^{\left(\sum_{i \epsilon Z} \delta_i \left(\frac{T_i - a_{in}}{b_{in}}\right) - \sum_{i \epsilon Z} \delta_i e^{\left(\frac{T_i - a_{in}}{b_{in}}\right)} - \sum_{i \epsilon Z} (1-\delta_i) e^{\left(\frac{L_i - a_{in}}{b_{in}}\right)} \right)} \\
&\quad \times e^{\left(\sum_{i \epsilon Z^c} \delta_i \left(\frac{T_i - a_{out}}{b_{out}}\right) - \sum_{i \epsilon Z^c} \delta_i e^{\left(\frac{T_i - a_{out}}{b_{out}}\right)} - \sum_{i \epsilon Z^c} \left(1-\delta_i\right) e^{\left(\frac{L_i - a_{out}}{b_{out}}\right)} \right)} \\
&= \left(b_{in}\right)^{-r_{in}} \left(b_{out}\right)^{-r_{out}} \\
&\quad \times e^{\left(\sum_{i \epsilon Z} \delta_i \left(\frac{t_i - a_{in}}{b_{in}}\right) - \sum_{i \epsilon Z} e^{\left(\frac{t_i - a_{in}}{b_{in}}\right)} \right)} \\
&\quad \times e^{\left(\sum_{i \epsilon Z^c} \delta_i \left(\frac{t_i - a_{out}}{b_{out}}\right) - \sum_{i \epsilon Z^c} e^{\left(\frac{t_i - a_{out}}{b_{out}}\right)} \right)}.
\end{aligned}
$$

Taking natural log on both sides,

$$\ln L(Z) = -r_{in} \ln b_{in} - r_{out} \ln b_{out} + \sum_{i \epsilon Z} \delta_i \left(\frac{t_i - a_{in}}{b_{in}} \right) - \sum_{i \epsilon Z} e^{\left(\frac{t_i - a_{in}}{b_{in}} \right)}$$
$$+ \sum_{i \epsilon Z^c} \delta_i \left(\frac{t_i - a_{out}}{b_{out}} \right) - \sum_{i \epsilon Z^c} e^{\left(\frac{t_i - a_{out}}{b_{out}} \right)}.$$

For $H_1 : b_{in} \neq b_{out}$ for at least one zone Z, the corresponding likelihood ratio statistic is

$$\lambda = \frac{\max_{Z, b_{in} \neq b_{out}} L(Z, b_{in}, b_{out})}{\max_{Z, b_{in} = b_{out}} L(Z, b_{in}, b_{out})} = \frac{L(\hat{Z})}{\hat{L}} \qquad (2.3.9)$$

where $\hat{Z}$ is the zone maximizing $L(Z, b_{in}, b_{out})$ under H_1, and $\hat{L}$ is the maximum of $L(Z, b_{in}, b_{out})$ under H_0.

The maximum likelihood estimators (MLE's) of the parameters $b_{in}, b_{out}, a_{in},$ and a_{out} for any arbitrary zone Z can be obtained by the following equations

$$\frac{\partial \ln L(Z)}{\partial b_{in}} = -\frac{r_{in}}{b_{in}} - \frac{1}{b_{in}^2} \sum_{i \epsilon Z} \delta_i \left(t_i - a_{in} \right) - \sum_{i \epsilon Z} e^{\left(\frac{t_i - a_{in}}{b_{in}} \right)} \left(\frac{-1}{b_{in}^2} \left(t_i - a_{in} \right) \right) = 0$$
$$\frac{\partial \ln L(Z)}{\partial b_{out}} = -\frac{r_{out}}{b_{out}} - \frac{1}{b_{out}^2} \sum_{i \epsilon Z^c} \delta_i \left(t_i - a_{out} \right) - \sum_{i \epsilon Z^c} e^{\left(\frac{t_i - a_{out}}{b_{out}} \right)} \left(\frac{-1}{b_{out}^2} \left(t_i - a_{out} \right) \right) = 0$$
$$\frac{\partial \ln L(Z)}{\partial a_{in}} = \frac{1}{b_{in}} \sum_{i \epsilon Z} (-\delta_i) - \sum_{i \epsilon Z} e^{\left(\frac{t_i - a_{in}}{b_{in}} \right)} \left(\frac{-1}{b_{in}} \right) = 0$$
$$\frac{\partial \ln L(Z)}{\partial a_{out}} = \frac{1}{b_{out}} \sum_{i \epsilon Z^c} (-\delta_i) - \sum_{i \epsilon Z^c} e^{\left(\frac{t_i - a_{out}}{b_{out}} \right)} \left(\frac{-1}{b_{out}} \right) = 0.$$

Thus the MLE's of the scale parameters b_{in} and b_{out} are

$\hat{b}_{in} = \frac{1}{r_{in}} \sum_{i \epsilon Z} (t_i - \hat{a}_{in}) \left[e^{\left(\dfrac{t_i - \hat{a}_{in}}{\hat{b}_{in}} \right)} - \delta_i \right]$ and

$\hat{b}_{out} = \frac{1}{r_{out}} \sum_{i \epsilon Z^c} (t_i - \hat{a}_{out}) \left[e^{\left(\dfrac{t_i - \hat{a}_{out}}{\hat{b}_{out}} \right)} - \delta_i \right]$, respectively.

Similarly, the MLE's of the location parameters a_{in} and a_{out} are obtained by the equations $r_{in} = \sum_{i \epsilon Z} e^{\left(\dfrac{t_i - \hat{a}_{in}}{\hat{b}_{in}} \right)}$ and $r_{out} = \sum_{i \epsilon Z^c} e^{\left(\dfrac{t_i - \hat{a}_{out}}{\hat{b}_{out}} \right)}$, respectively.

Under $H_1 : b_{in} \neq b_{out}$, the obtained MLE's provide

$$\begin{aligned} L(\hat{Z}) &= \left(\hat{b}_{in}\right)^{-r_{in}} \left(\hat{b}_{out}\right)^{-r_{out}} \\ &\quad \times e^{\left(\sum_{i \epsilon Z} \delta_i \left(\dfrac{t_i - \hat{a}_{in}}{\hat{b}_{in}} \right) + \sum_{i \epsilon Z^c} \delta_i \left(\dfrac{t_i - \hat{a}_{out}}{\hat{b}_{out}} \right) \right)} e^{(-r_{in} - r_{out})} \\ &= \left(\hat{b}_{in}\right)^{-r_{in}} \left(\hat{b}_{out}\right)^{-r_{out}} \\ &\quad \times e^{\left(\sum_{i \epsilon Z} \delta_i \left(\dfrac{t_i - \hat{a}_{in}}{\hat{b}_{in}} \right) + \sum_{i \epsilon Z^c} \delta_i \left(\dfrac{t_i - \hat{a}_{out}}{\hat{b}_{out}} \right) \right)} e^{-R}. \end{aligned}$$

Similarly, under $H_0 : b_{in} = b_{out}$,

$$\hat{L} = \left(\hat{b}_G\right)^{-r_{in}} \left(\hat{b}_G\right)^{-r_{out}} e^{\left(\sum_{i \epsilon G} \delta_i \left(\dfrac{t_i - \hat{a}_G}{\hat{b}_G} \right) \right)} e^{-R}$$

$$= \left(\hat{b}_G\right)^{-R} e^{\left(\sum_{i \epsilon G} \delta_i \left(\dfrac{t_i - \hat{a}_G}{\hat{b}_G} \right) \right)} e^{-R}.$$

So, the likelihood ratio statistic for $H_1 : b_{in} \neq b_{out}$ is

$$\lambda = \frac{\max_z \left(\hat{b}_{in}\right)^{-r_{in}} \left(\hat{b}_{out}\right)^{-r_{out}} e^{\left(\sum_{i\epsilon Z} \delta_i \left(\frac{t_i - \hat{a}_{in}}{\hat{b}_{in}}\right) + \sum_{i\epsilon Z^c} \delta_i \left(\frac{t_i - \hat{a}_{out}}{\hat{b}_{out}}\right)\right)}}{\left(\hat{b}_G\right)^{-R} e^{\left(\sum_{i\epsilon G} \delta_i \left(\frac{t_i - \hat{a}_G}{\hat{b}_G}\right)\right)}}.$$

In order to address the alternative hypothesis $b_{in} < b_{out}$, the function λ is multiplied by $I\left(\hat{b}_{in} < \hat{b}_{out}\right)$. Similarly, the function λ is multiplied by $I\left(\hat{b}_{in} > \hat{b}_{out}\right)$ if the alternative hypothesis $b_{in} > b_{out}$ is under consideration.

2.4 Permutation Test Procedure

Since there is no closed analytical form of the distribution of the test statistic λ, the standard analytical p-values cannot be calculated. Instead a permutation test procedure was used to test the statistical inference of the selected clusters. Unlike most of the scan statistics, it is not possible to generate the simulated data under the null hypothesis, since the distribution of the observed time to events is unknown. To overcome this situation, the observed time to event data along with its corresponding censoring indicators were permuted a large number of times and the statistical significance was calculated. This permutation step ensures that no matter how the observed time to events data is distributed, this distribution is preserved for each permuted dataset. This factor provides valid statistical significance since all the permuted datasets are equally distributed, irrespective of their spatial locations.[18]

In particular, the observed pairs $\{(t_i, \delta_i)\ i = 1, 2, \ldots, N\}$ were permuted among the individual geographical coordinates of the original study region.[18] For each permuted dataset, the log-likelihood was calculated for each zone and the most likely cluster preserving the maximum log-likelihood in the dataset was saved.

In the permutation test procedure, a p-value is calculated as the fraction of permutations that are atleast as extreme as the test statistic from the observed time to event data.[38] No matter how the time to event data are distributed, the SSS using the Weibull distribution does not provide biased p-values associated with the potential clusters,[19] and is also true for the SSS's based on the exponential and log-Weibull distributions. This feature occurs because the randomly permuted time to event data along with its permuted co-ordinates preserve the correct α-level for any data distribution.[18,19]

Secondary clusters are the spatial clusters that do not overlap with the most likely cluster, and reject the null hypothesis of no clustering on their own.[18] These clusters are ranked with their corresponding LLR values and the associated p-values are calculated in the same spirit of the main cluster by comparing the k^{th} (say) highest likelihood in the real dataset with the maximum likelihood in the randomly permuted datasets.[18] All of the significant secondary clusters were also reported with the most likely cluster in this study.

2.5 Summary

The SSS is a statistical technique for identifying the presence and location of the significant spatial clusters with certain elevated characteristics. In this chapter, a newly established SSS for time to event data based on the log-Weibull distribution has been developed. This SSS is capable of using both, the censored and uncensored observations. The test statistic is based on the log-likelihood ratio and is tested using the permutation testing approach. Already developed SSS's for the exponential and Weibull distributions have also been defined briefly.

Chapter 3

Data Description and Analysis

3.1 Overview

Administrative data from Alberta Health are described in Section 3.2. Section 3.3 contains the results for spatial scan statistics (SSS's) using the exponential, Weibull, and log-Weibull distributions. A summary of the spatial scan results follows in Section 3.4.

3.2 Data Description

3.2.1 Study Design

This retrospective cohort study used the data for patients discharged from the emergency department (ED) who presented with atrial fibrilation and flutter (AFF) in the province of Alberta during a 1-year period: April 1, 2010, to March 31, 2011.

3.2.2 Study Population

Alberta, with a land area of 661,848 km^2 and a population size of 3,645,257 in 2011,[40] is the fourth largest province of Canada.[41] Under the Canada Health Act of 1984, all Alberta residents have access to a publicly administered and funded health care system.[42] This system ensures that all Albertans have universal access to health care services.

In 2003, the province of Alberta was divided into nine administrative health areas also called Regional Health Authorities (RHA's).[43] These RHA's were further partitioned into 70 sub-Regional Health Authorities (sRHA's) and are shown in Appendix-A (Table A.1) and Figure 3.1. In 2008, the nine RHA's were combined to form Alberta Health Services and it was formally launched on April 1, 2009.[44] Alberta Health Services was further organized into five geographic zones (South, Calgary, Central, Edmonton, and North), each with different population sizes and geographical boundaries.[45] These health zones provide ease for the administration of health services locally.

The sRHAs have diverse population sizes ranging from 550 to 140,211 with a median population size of 46,075 in 2011. Each sRHA was considered as a geographical unit for the analysis. For each sRHA's centroid based on population, the latitude and longitude of the centroids were developed by Alberta Health.[43] Distances between the pairs of sRHA population-based centroids were ordered and used to create the nearest neighbours, shown in Appendix-A (Table A.2).

Emergency departments are central and important units in the Alberta health care system for medically ill and injured patients to get around the clock emergency care. For example, for patients with sudden cardiac conditions, ED's may be used for immediate care. Hospitals located in urban and rural areas are equipped with emergency physicians on a fulltime or on-call availability basis depending upon the volume of patients.[46] Albertans receive emergency care in more than 100 public funded hospitals.

Atrial fibrillation and flutter is one of the most common clinically diagnosed forms of arrhythmia frequently observed in outpatient settings.[47] It is caused when the atria beats abnormally fast because of the development of an irregular conduction circuit inside the right atrium.[48] This leads to an unbalanced heart rhythm and poor blood flow. In the last 20 years, a 60% increase in hospital admissions for AFF has been noticed, the major reasons for which are the aging population and an increase in chronic heart disease.[49,50]

Figure 3.1: Alberta sub-Regional Health Authorities

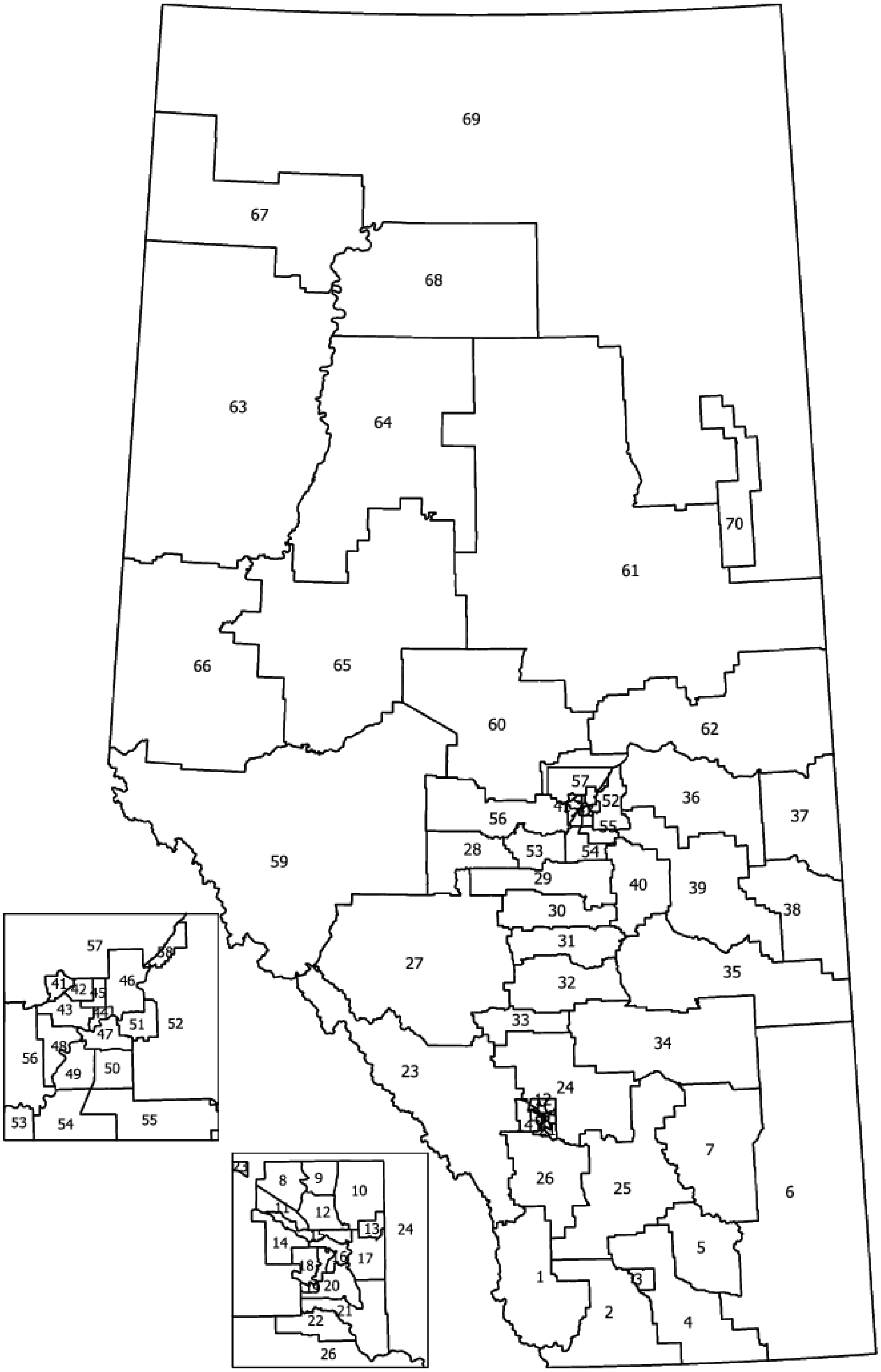


Table 3.1: Codes and corresponding Regional Health Authorities in Alberta

Codes	Regional Health Authority (RHA)
1 – 5	Chinook Regional Health Authority (R1)
6 – 7	Palliser Health Region (R2)
8 – 26	Calgary Health Region (R3)
27 – 35	David Thompson Regional Health Authority (R4)
36 – 40	East Central Health (R5)
41 – 58	Capital Health (R6)
59 – 62	Aspen Regional Health Authority (R7)
63 – 66	Peace Country Health (R8)
67 – 70	Northern Lights Health Region (R9)

3.2.3 Data Sources and Description

As a part of the activities of Alberta Health, administrative health datasets are regularly maintained and are available for planning and research purposes. There were three main data sources used in this study: the Ambulatory Care Classification System, the Alberta Health Care Insurance Plan, and the Physician's Claim File.

Ambulatory Care Classification System (ACCS)

The ACCS database was developed by Alberta Health and put into practice in 1997 to track the ambulatory care visits within government funded hospitals.[51] An ambulatory care service is defined as an approach to a regional health service provider by a patient without the condition of staying at the hospital.[51] One of the uses of the collected ACCS data is the classification of patients receiving ambulatory care services into clinical groups according to their clinical profiles and required resources.[52]

The ACCS database includes ED visits, clinic visits, and outpatient services delivered within acute care institutions in Alberta,[51] which are used for the analysis and management purposes at both the hospital and provincial levels.[52] Information regarding the patient's identification number, ED presentation timings and dates, disposition status, and diagnosis information are entered into computerized abstracts.[51] All patients exiting the ED are allocated a disposition status. Only ED visits ending in discharge were considered in this study. If the patient had multiple ED visits, then the last visit was considered.

The National Ambulatory Care Reporting System (NACRS) based on ACCS data system and commenced in 1997, is an effective database managed by Canadian Institute for Health Information (CIHI).[53,54] The NACRS provides standardized data collection tools and reporting techniques to help the health care institutions and community-based organizations to keep record of the ambulatory care visits.[54]

Alberta Health Care Insurance Plan (AHCIP)

This data source contains the demographic information (e.g., age, gender, and health sub-region of residence) for all people registered under the provincial health insurance plan.[55] The required information for this study was obtained by linking the discharged ED patients in the ACCS database to the AHCIP registry file. In addition, the AHCIP provided data for the population by the demographic variables.

Physician's Claim File (PCF)

The Physician's Claim File (PCF) was used to obtain the information for patients' visits in non-ED settings (follow-up visits). This information was collected for all the patients who had ED presentations for AFF that ended in discharge during 365 days of study period (i.e., April 1, 2010, to March 31, 2011). The required datasets were obtained by linking the patients in the ACCS to the patients in the Physician's Claims File database. The obtained follow-up visits data were comprised of the date of the physician visit and the speciality of the physician. A physician specialist visit in this study was defined as either cardiology (CARD) or internal medicine (INMD). If a patient had both cardiology and internal medicine follow-up visits, the first visited physician was considered (i.e., earlier in date and time). Table 3.2 provides a list of the data fields and sources used in this study.

Table 3.2: Data sources used in the study

Variable	**Source**
Diagnostic Information for ED Visit	ACCS
Disposition Status for ED visit	ACCS
Start/End Dates/Time of ED visit	ACCS
Age	AHCIP
Gender	AHCIP
Health Region of Residence	AHCIP
Date of Follow-up Visit	PCF
Diagnostic Information for Follow-up Visit	PCF
Physician Specialty at Follow-up Visit	PCF

3.2.4 Case Definition and Outcome of Interest

The key outcome of interest was the time from ED discharge for AFF to the 1^{st} specialist visit during 365 days of the study period, i.e., April 1, 2010, to March 31, 2011. The patient could have a specialist follow-up visit in the time span of ED end time, to the end of the study. Each discharged ED presentation during April 1, 2010, to March 31, 2011 with a follow-up visit to the specialist during ED end time, to March 31, 2011 was considered as a complete time to event outcome. If the patient had no specialist visit by the end of the study (March 31, 2011), the outcome was referred to as right (type-I) censored. If the patient had both ED discharge and the follow-up visit on the same day, the time to event was rounded to a whole day. Each Alberta resident making at least one ED presentation for AFF during the fiscal year was referred to as a case (patient).

The methodology used in this study does not adjust for repeated ED presentations of cases. Hence, independent patient data was considered by taking only the last ED visit out of the multiple visits, if they occurred. This study focused on detecting spatial clusters of patients in Alberta having longer times to follow-up visits to cardiology or internal medicine specialists, who were discharged from ED. The calculations were performed using the R and S-Plus softwares.[56,57] With the SSS, each detected cluster could contain only a maximum of $r^* = 10\%$ (a pre-decided upper bound) of the total population, so the variable scanning windows were created around each sRHA to absorb neighbours up to a fixed 10% of the total population. This upper bound was chosen based on the feasibility of analysis and time restrictions.

3.2.5 Exploratory Data Analysis

During the study period, there were about 1.95M adults in the population, aged 35 years and above, with an average age of 54.11 years. The reason behind considering the age $\geq$ 35 in this study was that AFF is more prevalent in older adults and covers more homogeneous groups.[58] In total, 5,953 ED cardiac visits were made during the study period, among them 4,292 were for the AFF. The discharged subset excluding the multiple ED presentations for AFF (if they existed) and including only the last ED visits, was comprised of 3,527 cases and the average age of these individuals was 68.15 years (Table 3.3). The median time taken by the specialists to see ED discharged patients for AFF was 151 days. Approximately 20% of the observations were censored. A set of (t_i, δ_i) was observed from $i = 1, ..., 3,527$ patients, where $t_i = \min(T_i, L_i)$ with T_i being the time to event and L_i representing the fixed right censoring time. The indicator of complete or censored observation is δ_i.

Summaries of the administrative data by sub-Regional Health Authorities were calculated for both the population and cases under study and are shown in Table 3.4. This table shows that during 2010-2011, on average there were approximately 28,000 people living in each sRHA. An average of 50 presented to an ED for AFF with a mean age of 68 years. An average of 10 censored outcomes were observed in each sRHA. Twenty seven out of 13,861 males were observed on average presenting to ED for AFF in each sRHA.

Table 3.3: Summary of the administrative data

	Population		Cases	
N	1,953,830		3,527	
Censored, N	–		687	(19.40%)
Male, N	970,338	(49.60%)	1,909	(54.00%)
Age				
Mean (SD)	54.11	(13.45)	68.15	(13.83)
Med (Min, Max)	57.01	(35, 112)	68.81	(35, 97)
ED presentations	–		5,953	
AFF cases	–		4,292	
Time to 1^{st} follow-up				
Med (Min, Max)	–		151	(1, 365)

Med=Median	Min=Minimum
Max=Maximum	SD=Standard Deviation

Table 3.4: Summary of the administrative data by sub-Regional Health Authorities

	Population		Cases	
N				
Mean (SD)	27,912	(18,154.46)	50	(30.86)
Med (Min, Max)	26,152	(2,573, 70,510)	45	(3, 123)
Censored, N				
Mean (SD)	–		10	(7.23)
Med (Min, Max)	–		8	(0, 35)
Males, N				
Mean (SD)	13,861	(8,895.56)	27	(16.22)
Med (Min, Max)	12,857	(1,282, 34,731)	24	(1, 73)
Age, Means				
Mean (SD)	54.39	(2.04)	68.03	(4.51)
Med (Min, Max)	54.40	(48, 58)	67.58	(53, 85)

Med=Median	Min=Minimum
Max=Maximum	SD=Standard Deviation

Figure 3.2: Total number of uncensored and censored observations per sub-Regional Health Authority. Uncen=Uncensored observations, Cen=Censored observations

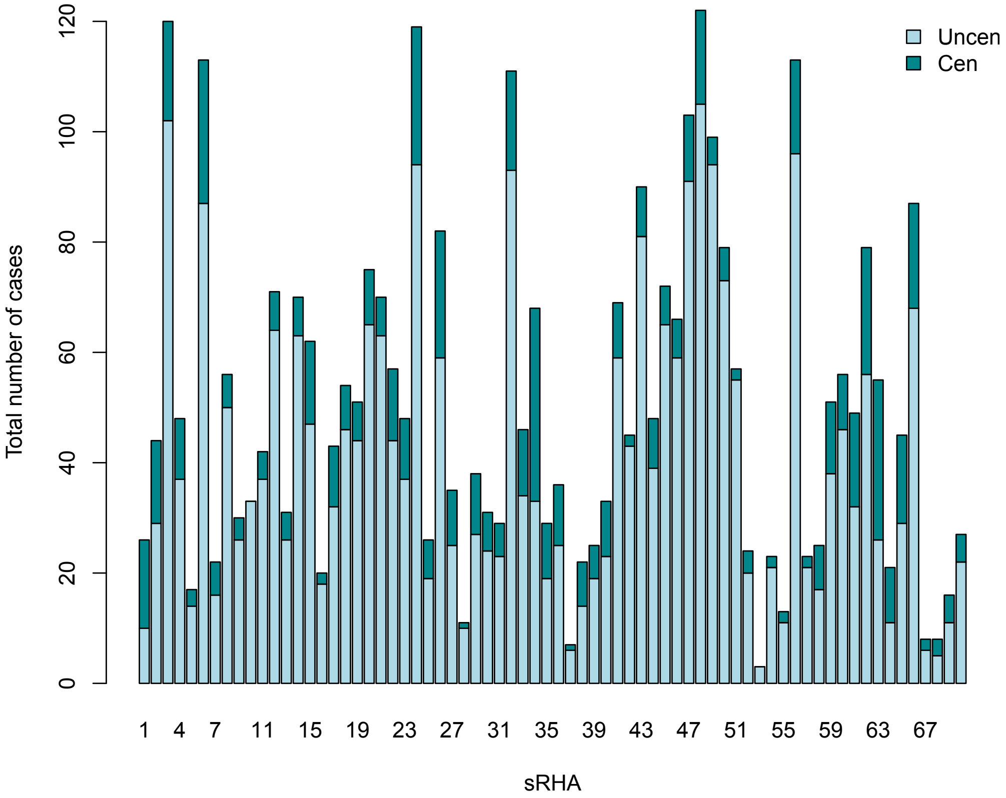
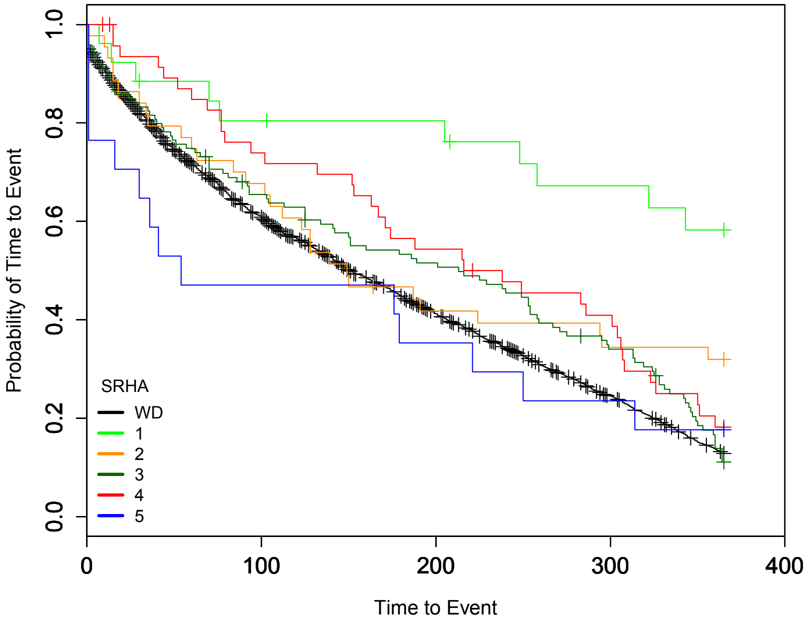


Figure 3.2 shows the total number of uncensored (complete) and censored time to events for each sRHA. There were a total of 3,527 cases who were discharged from the ED for AFF. Among these cases, 2,840 provided uncensored information and 687 were censored outcomes. These numbers are also presented in Table A.3 (Appendix A).

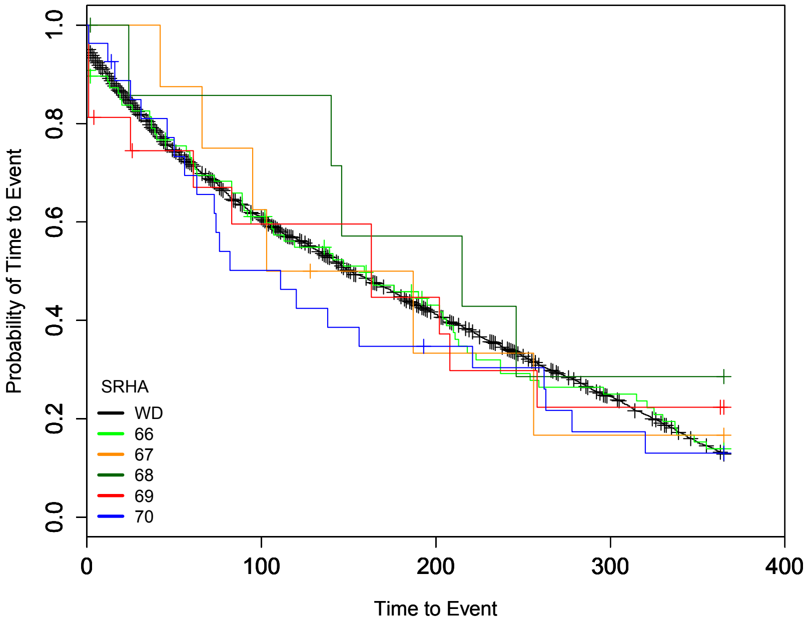
Kaplan Meier (KM) plots adjusted for censoring, for the whole data and a few sRHAs, are shown in Figure 3.3 to show the estimated probabilities of times to event. The plots show variation in the time to 1^{st} specialist visit by sRHA. The median time to event for the whole dataset was 151 days and the corresponding 95% lower and upper confidence limits were 144 to 161 days, respectively.

Figure 3.4 shows the KM estimated median (the black dot in each bar) and lower and upper quartiles (lower and upper borders of each bar) for each sRHA. These estimated values were calculated after adjusting for the censored information. The figure depicts the variation by sRHAs in the study region and that the majority of the sRHAs had median time to events between 100 and 200 days. Among all 70 sRHAs used in the study, the maximum median time to event (364 days) was observed for R63 and the minimum was for R5 (54 days). A solid horizontal line represents the overall KM estimated median for time to specialist visits for 3,527 discharged cases, with the corresponding lower and upper quartiles provided as dotted lines.

Figure 3.3: Kaplan Meier plots for the whole data and selected sub-Regional Health Authorities

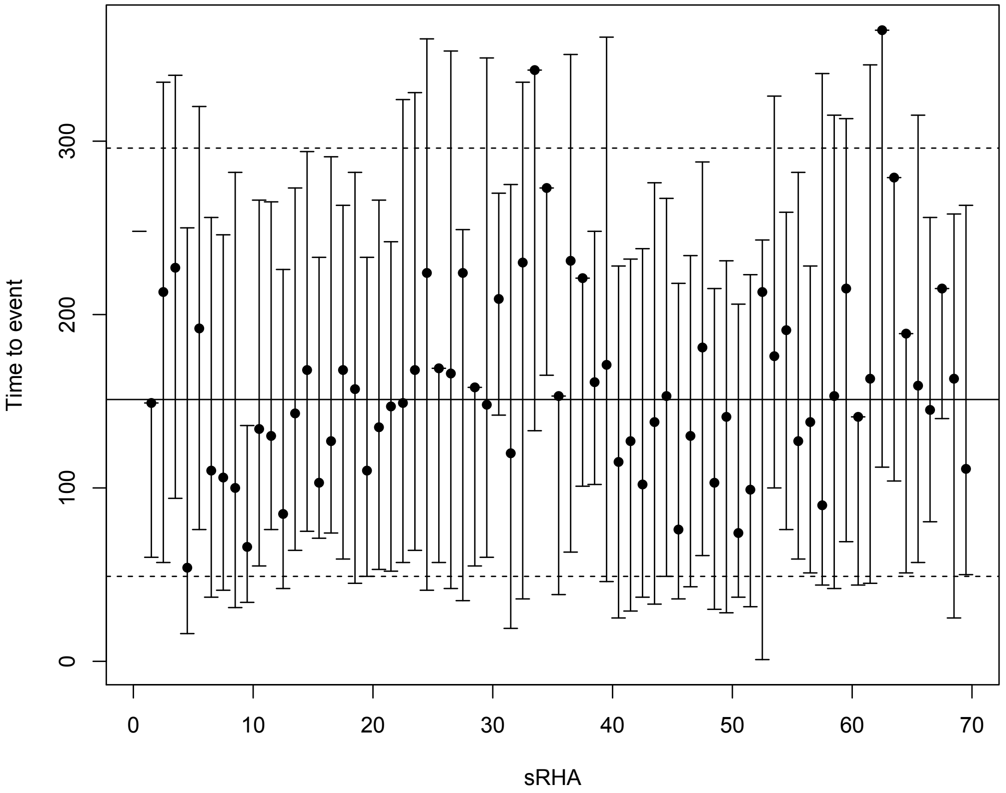


(a) Kaplan-Meier plots for whole data (WD) and sRHAs 1-5



(b) Kaplan-Meier plots for whole data (WD) and sRHAs 66-70

Figure 3.4: KM estimated median and lower and upper quartiles for each sub-Regional Health Authority



3.3 Analysis and Results from the Spatial Scan Statistics

In this section, the candidate main clusters along with the statistically significant secondary clusters, based on three probability distributions (i.e., the exponential, Weibull and log-Weibull) are shown. Each method identified geographical areas with longer than expected times to a specialist follow-up visit.

3.3.1 Exponential Spatial Scan Statistic

The identified main and secondary clusters for longer time to specialist visits for patients discharged from the ED who presented with AFF using the exponential distribution are shown in Table 3.5. Highlighted detected main and secondary clusters are also shown on the Alberta map (Figure 3.6) in red and orange colours, respectively.

Table 3.5 shows that the most likely cluster with significantly longer times to events was mainly from the Peace Country Health region, Northern Lights Health Region, and the Aspen Regional Health Authority. This cluster was identified with 202 observed number of cases. The log-likelihood ratio (LLR) was 35.87 with the associated p-value (P) of 0.004. The detected secondary cluster with the p-value of 0.013 was mainly at the hub of the East Central Health Authority, and parts of the David Thompson RHA and the Capital Health Authority.

Table 3.5: Spatial scan results for the exponential distribution

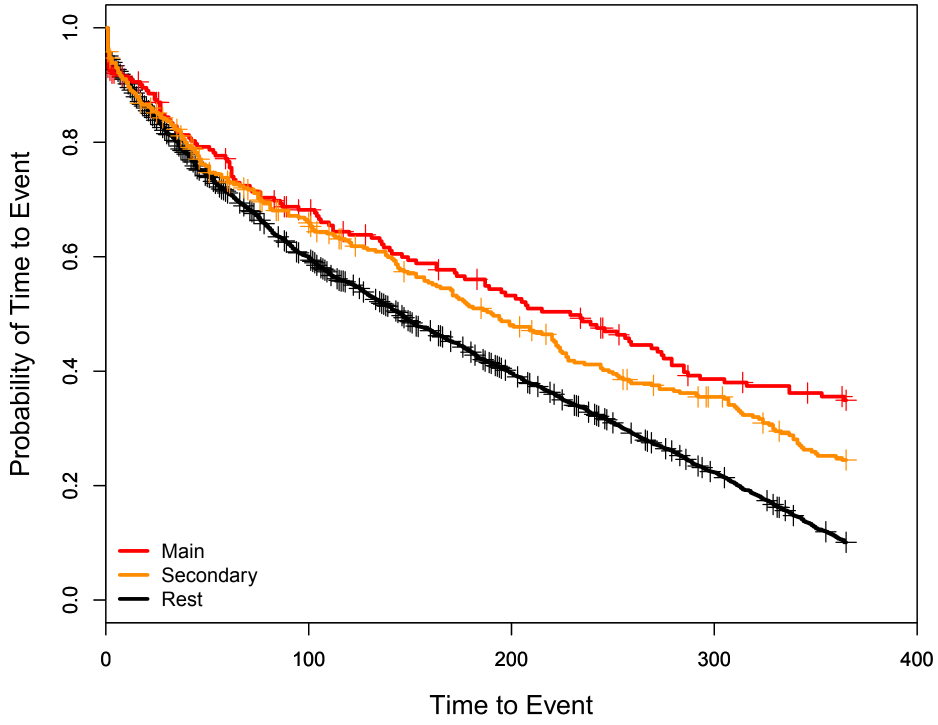
Cluster	Pop	Cases	LLR	P	M-In	M-Out
Main 68 64 67 63 69 65 61	61,277	202	35.87	0.004	229	148
Secondary 38 37 35 39 36 40 62 34 52 55 58	154,011	360	32.17	0.013	190	148

Pop=Population
LLR=Log-Likelihood Ratio
M-In=Medain time to 1^{st} specialist visit within the cluster
M-Out=Medain time to 1^{st} specialist visit outside the cluster

The median times to 1^{st} follow-up visits for inside and outside the detected main and secondary clusters were also obtained by accounting for the censored information. Within the main and secondary clusters, the medians were 229 and 190 days, respectively. The corresponding 95% confidence intervals (CI's) were (173, 279) for inside the main cluster and (163, 223) for inside the secondary cluster, respectively. Outside the main and secondary clusters, the median was 148 days. That is, for the entire province without the main cluster, the median was 148 days with (141, 159) being the 95% CI. In addition, for the entire province without the secondary cluster, the median was also 148 days with 141 to 158 days as the confidence bounds. The median event time for the whole province excluding both the main and secondary clusters was 145 days, with (137, 153) as the 95% CI.

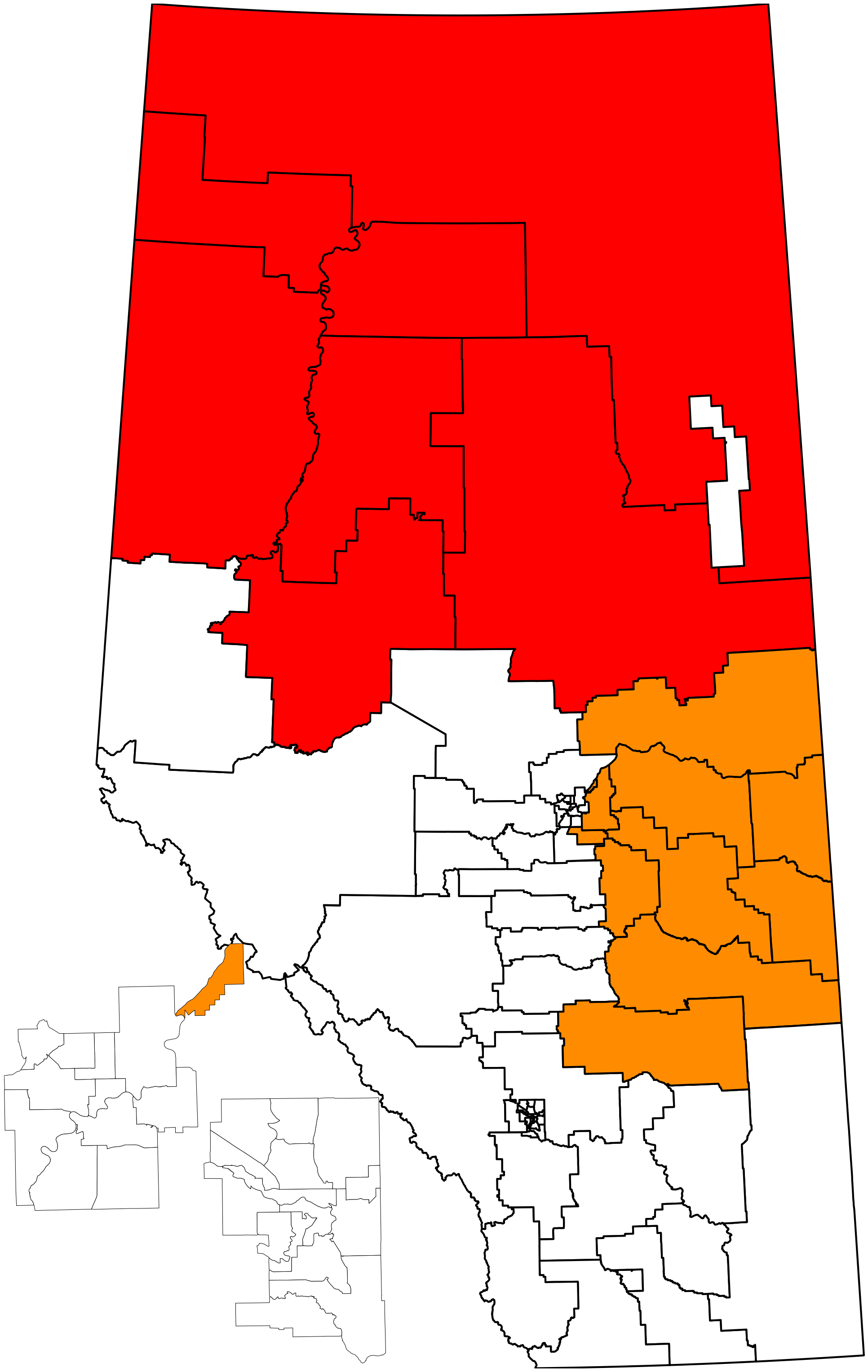
Figure 3.5 shows the KM curves for the times to event for the three distinct areas: the main cluster, the secondary cluster, and rest of the province (exclusive of the main and secondary clusters). Since we were detecting the clusters with

Figure 3.5: Kaplan Meier curves for the detected main and secondary clusters and rest of the province for time to 1^{st} specialist visit for the exponential spatial scan statistic



longer times to specialist visits, the estimated KM curve for the main cluster showed the highest probabilities of seeing a specialist near 365 days.

Figure 3.6: Alberta Map highlighting the main and secondary clusters for the exponential Distribution. (Red=Main cluster, Orange=Secondary cluster)



3.3.2 Weibull Spatial Scan Statistic

The identified main and secondary clusters are highlighted on the Alberta map (Figure 3.8) with red and orange colours, respectively, and the results are presented in Table 3.6.

The Weibull SSS detected the same most likely cluster as of the exponential SSS, i.e., the combination of the Peace Country Health region, Northern Lights Health Region, and the Aspen Regional Health Authority. The corresponding value of the LLR was 95.22 with the associated p-value of 0.001. The detected secondary cluster showed a narrower zone than the exponential SSS's secondary cluster, which had an LLR= 35.89 and p-value= 0.001. The major geographical areas detected in the secondary cluster were the East Central Health Region with some parts of the David Thompson RHA and the Capital Health Authority.

Table 3.6: Spatial scan results for the Weibull distribution

Cluster	**Pop**	**Cases**	**LLR**	**P**	**M-In**	**M-Out**
Main 68 64 67 63 69 65 61	61,277	202	95.22	0.001	229	148
Secondary 38 37 35 39 36 40 62 34 52	136,554	323	35.89	0.001	194	147

Pop=Population
LLR=Log-Likelihood Ratio
M-In=Medain time to 1^{st} specialist visit within the cluster
M-Out=Medain time to 1^{st} specialist visit outside the cluster

Median times to 1^{st} specialist within the main and secondary clusters were 229 and 194 days, respectively. The associated 95% CI's were (173, 279) for the main cluster and (170, 225) for the secondary clusters, respectively. Similarly, the median times were 148 and 147 days for the entire province excluding the main and secondary cluster, respectively. The CI's were (141, 159) for the area excluding the main cluster and (141, 157) for the area without the secondary clusters. By excluding both the main and secondary clusters, the median time to events for the rest of the province was calculated as 145 days and the corresponding lower and upper confidence bounds were 137 to 153 days, respectively. It is worth noting that the observed time to events data (complete and censored) used in study follows the Weibull distribution with the estimated shape and scale parameters be 0.884 and 194.356, respectively.

The KM curves for the main and secondary clusters and the rest of the province detected by the Weibull SSS were created. As expected, the KM curve for the main cluster was above the other curves.

Figure 3.7: Kaplan Meier curves for the detected main and secondary clusters and rest of the province for time to 1^{st} specialist visit for the Weibull spatial scan statistic

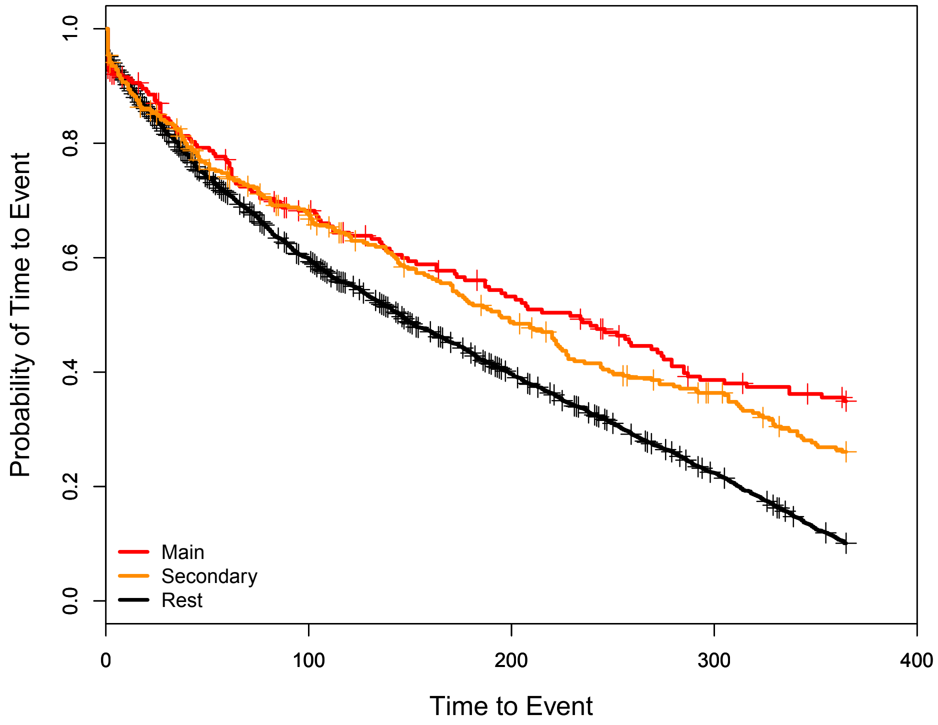
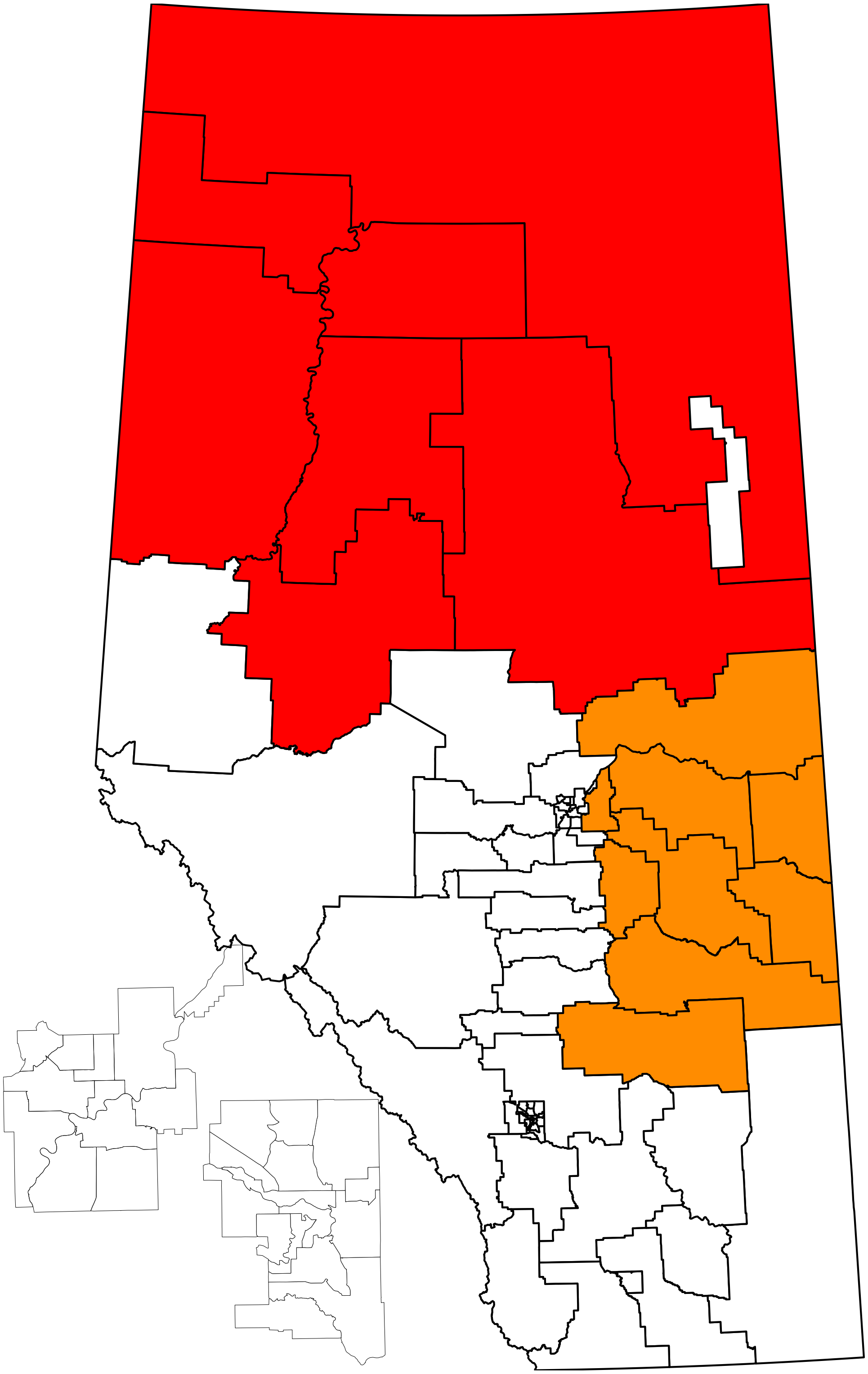


Figure 3.8: Alberta Map highlighting the main and secondary clusters for the Weibull Distribution. (Red=Main cluster, Orange=Secondary cluster)



3.3.3 Log-Weibull Spatial Scan Statistic

The log-Weibull SSS as a new method was used to detect the main and secondary clusters and the results are shown in Table 3.7. Highlighted main and secondary clusters are presented in Figure 3.10. The results for the log-Weibull SSS were not much different from the exponential and Weibull SSS's.

The SSS with the log-Weibull distribution detected the same most likely cluster as of the exponential and Weibull SSS with LLR = 76.23 and p-value = 0.001. This SSS provided two different statistically significant secondary clusters, one of them was a narrower form of the already detected secondary cluster by the exponential and Weibull distributions, i.e., a combination of the David Thompson RHA and the East Central Health Authority. The other one was a part of the Chinook Regional Health Authority.

Table 3.7: Spatial scan results for the log-Weibull distribution

Cluster	**Pop**	**Cases**	**LLR**	**P**	**M-In**	**M-Out**
Main 68 64 67 63 69 65 61	61,277	202	76.32	0.001	229	148
Secondary(1) 38 37 35 39 36 40 62 34	121,880	299	70.93	0.001	199	147
Secondary(2) 1 2	20,238	71	17.83	0.032	292	150

Pop=Population
LLR=Log-Likelihood Ratio
M-In=Medain time to 1^{st} specialist visit within the cluster
M-Out=Medain time to 1^{st} specialist visit outside the cluster

Median times to event were 229, 199, and 292 days for inside the main, secondary (1), and secondary (2) detected clusters, respectively. The corresponding 95% lower and upper confidence bounds were, 173 to 279, 171 to 231, and 150 to 295 days. For the entire province, collectively excluding the main and both secondary clusters, the median event time was 144 days and the 95% CI was (135, 152) days, respectively. Figure 3.9 shows the KM curves for the detected main and secondary clusters and the rest of the province. The KM curve for the secondary (2) cluster is on a slightly higher position than the main and the secondary (1) clusters. One of the reasons can be the less number of patients and higher percentage of censoring in secondary (2) cluster, which moved its KM curve higher than the others.

Figure 3.9: Kaplan Meier curves for the detected main and secondary clusters and rest of the province for time to 1^{st} specialist visit for the log-Weibull spatial scan statistic

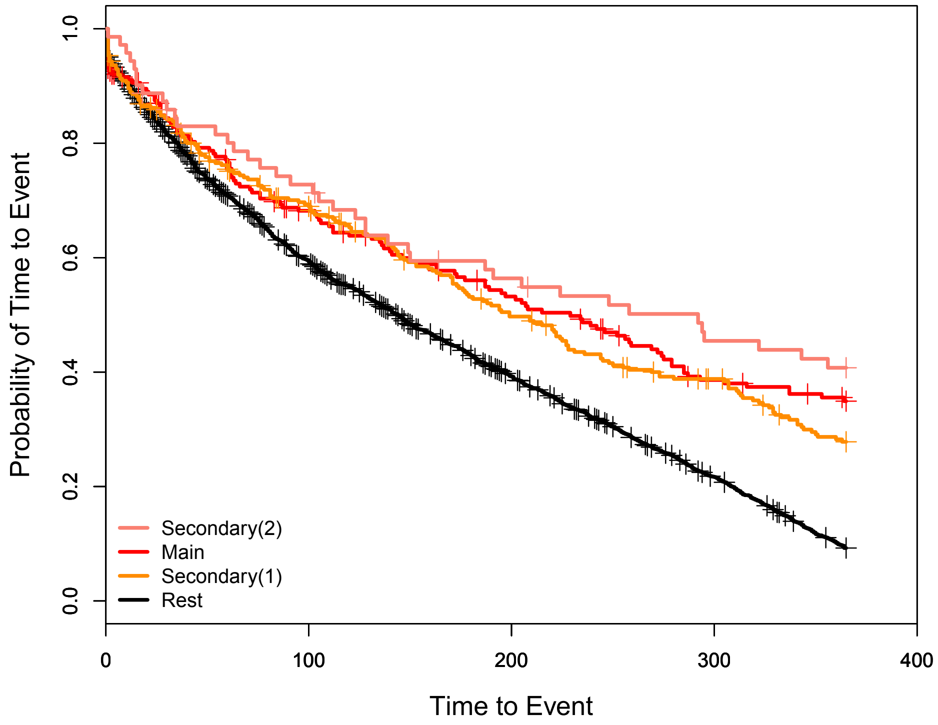
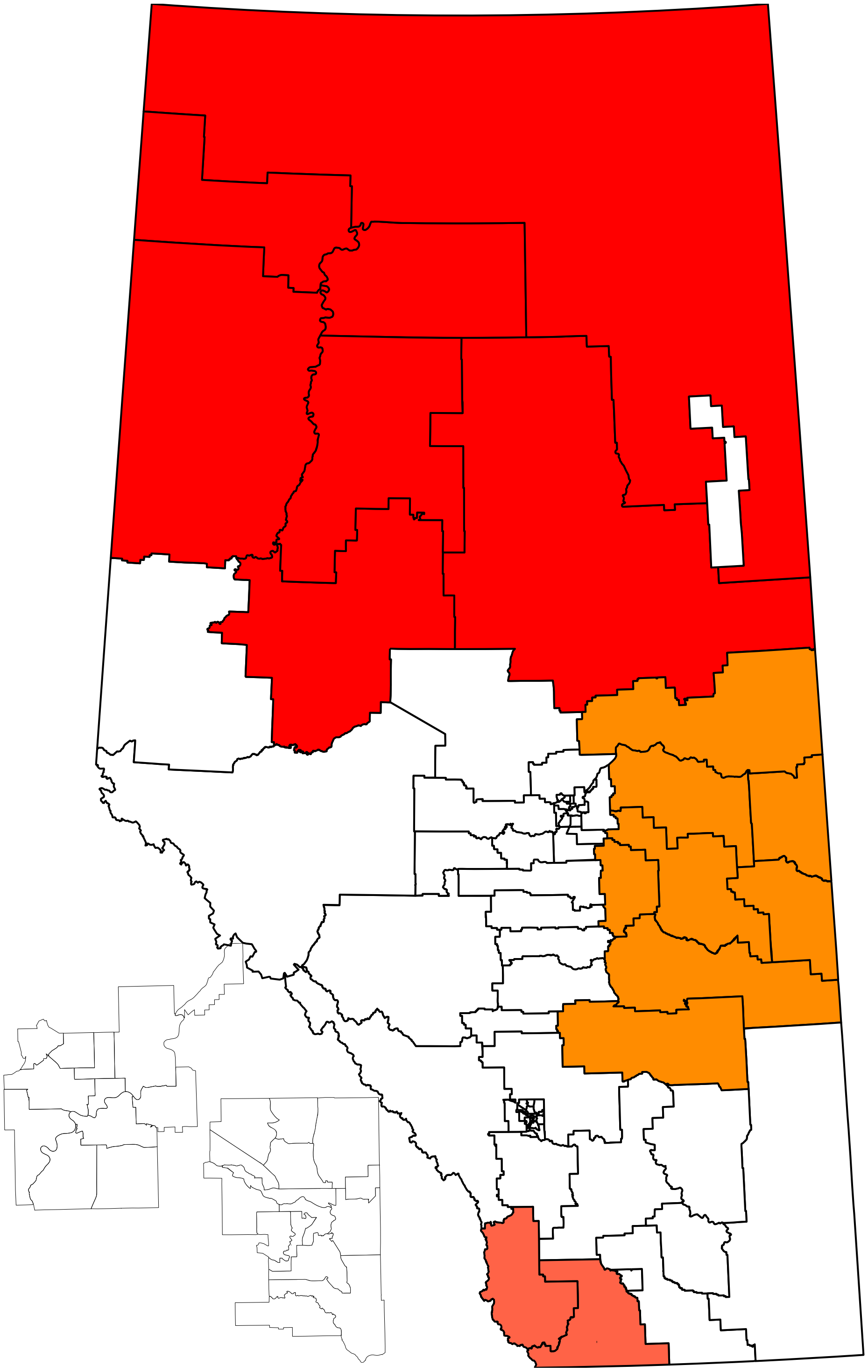


Figure 3.10: Alberta Map highlighting the main and secondary clusters for the log-Weibull Distribution. (Red=Main cluster, Orange=Secondary cluster (1), Dark orange = Secondary cluster (2))



3.4 Summary

A summary of the spatial scan results for all three distributions is presented in Table 3.8. Each SSS approach detected the same main cluster of longer times to specialist visit after an ED presentation for AFF. Similarly, the sRHAs detected under the secondary cluster (1) by each SSS were approximately the same, with the exponential SSS providing the widest geographical area. The log-Weibull SSS was able to detect one more statistically significant secondary cluster with a p-value of 0.041.

Table 3.8: Summary of the results for the spatial scan statistics

Cluster	**Distribution of the SSS**		
	Exponential	**Weibull**	**Log-Weibull**
Main			
sRHA's	68 64 67 63 69 65 61	Same as Exponential	Same as Exponential
LLR	35.875	95.215	76.32
P-value	0.004	0.001	0.001
Sec(1)			
sRHA's	38 37 35 39 36 40 62 34 52 55 58	38 37 35 39 36 40 62 34 52	38 37 35 39 36 40 62 34
LLR	32.172	35.896	70.93
P-value	0.013	0.001	0.001
Sec(2)			
sRHA's	–	–	1 2
LLR	–	–	17.83
P-value	–	–	0.032

LLR=Log-Likelihood Ratio
Sec=Secondary

Chapter 4

Simulation Studies

4.1 Overview

The simulation study procedures are described in Section 4.2. The exponential, Weibull, and log-Weibull spatial scan statistics (SSS's) under four different probability models for right differential censoring are compared and contrasted in Section 4.3. The summary of the chapter follows in Section 4.4.

4.2 Strategy Used for Simulation Studies

Simulation studies were conducted to investigate the power of detecting a potential cluster, the effect of right differential censoring on cluster detection, and the strength of detection of a true cluster using simulated datasets from the exponential (Exp), Weibull (Weib), log-Normal (LN), and gamma (Gam) probability models. All of the datasets were analyzed with the exponential, Weibull, and log-Weibull SSS's.

To perform the simulation studies, time to event data were randomly generated for 500 individuals (cases). To test the behaviour of the SSS's, datasets were generated from four different probability models, i.e., the exponential, Weibull, log-Normal, and gamma under various parametric situations leading to diverse means and variances inside and outside the zone designated as the true cluster. The Alberta geography was used as the geography for analysis and the Alberta population was used to create the zones for the simulation studies. Like the spatial scan analysis of the real administrative data, an upper bound of 10% was imposed on the population size under the continuously increasing searching windows because of the time restrictions and feasibility of the simulation study analyses.

For all simulated datasets, a true cluster of 25 individuals was created at the Palliser North and Central Health Region (ID=6), to have longer time to events than the rest of the province. Right differential censoring was added with the ratios of 20%:20%, 20%:40%, and 40%:20% for inside:outside the true cluster. More precisely, three different scenarios of right censoring were incorporated. In the first case, 20% censoring was used both within and outside the true cluster. In the second case, 20% censoring was used within the true cluster and 40% outside the true cluster, whereas this ratio was reversed in the third case. One of the purposes behind this simulation study was to examine the behaviour of the SSS's under various censoring situations.

One thousand simulated datasets were generated from the exponential, Weibull, log-Normal, and gamma probability models using the differential censoring settings described above under the alternative hypotheses of the existence of longer than expected time to event clusters. For symmetry, parameters for each probability model were chosen in such a way that they provided a constant mean of 2 outside the true cluster and means of 10, 15, and 20, inside the true cluster with corresponding variances. There were three time to event datasets' means inside the true cluster for each probability model under all censoring ratios. The variances for the data outside the true cluster were 4, 0.188, 2, and 1 for the exponential, Weibull, log-Normal, and gamma models, respectively.

For each simulated dataset, 999 random permutations were performed to get the p-values from the permutation testing procedure. Power (Pow) was calculated as the proportion of datasets out of 1000 having p-values< 0.05,[19] not necessarily detecting the true cluster. In order to observe the strength of identification of the true cluster by each SSS, three different proportions were calculated for mutually exclusive situations from 1000 randomly generated datasets under each probability model for all censoring situations. These were the proportion of datasets:

1. Perfectly identifying the true cluster (PI);
2. Identifying a large cluster including the true cluster (LC); and,
3. Not identifying the true cluster (NI).

Mathematically, we define

Z^* = True cluster.

$Z^{(m)}$ = The cluster identified in m^{th} simulation.

M = Total simulations.

The key quantities were defined as,

$$\text{Power} = \frac{1}{M}\sum_{m=1}^{M} I_{[Z^{(m)}\ ;\ P(Z^{(m)})<\ 0.05]} \tag{4.2.1}$$

$$\text{PI} = \frac{1}{M}\sum_{m=1}^{M} I_{[Z^{*}=Z^{(m)}]} \tag{4.2.2}$$

$$\text{LC} = \frac{1}{M}\sum_{m=1}^{M} I_{[Z^{*}\subset Z^{(m)}]} \tag{4.2.3}$$

$$\begin{aligned}\text{NI} &= \frac{1}{M}\sum_{m=1}^{M} I_{[Z^{*}\not\subseteq Z^{(m)}]} \qquad (4.2.4)\\ &= 1 - \text{PI} - \text{LC}.\end{aligned}$$

4.3 Simulation Study Results

4.3.1 Simulation Study for Exponential Spatial Scan Statistic

The results from the simulation study for the exponential SSS are provided in Table 4.1, and also depicted in pictorial form in Appendix B (Figures B.1 and B.2).

Table 4.1 contains the power values of detecting the potential clusters, the proportions of datasets perfectly identifying a true cluster (PI), and the proportions of datasets identifying large clusters including the true cluster (LC).

For the 20%:20% censoring setting, the power values of detecting a potential cluster were between 0.705 and 1 for all of the probability models used in the study. The values of the proportions of the datasets, that detected a perfect true cluster were between 0.554 and 1. No dataset was captured with a non-zero proportion of not identifying a true cluster.

Under the 20%:40% censoring ratio situation, the powers ranged from 0.645 to 1 for all of the probability models. The values of the proportions of datasets detecting a perfect true cluster were between 0.545 and 1, being highest for the exponential and lowest for the log-Normal simulated time to event datasets. The proportions of the datasets not identifying a true cluster were all zeroes under the 20%:40% censoring ratio.

For the 40%:20% right censoring case, the power values were between 0.403 and 0.941 and the proportions of detection of a perfect true cluster ranged from 0.119 to 0.957. The datasets generated under the log-Normal probability model had the highest proportions of identifying large clusters including the true cluster. There were also a few situations of not detecting a true cluster at all in this censoring ratio, when the simulated data were generated from the Weibull and gamma models.

Overall, a visible decrease was seen in the power of detection of a potential cluster and the proportions of datasets identifying a perfect true cluster from the 20%:20% to 20%:40% scenarios and the 20%:40% to 40%:20% scenarios under each probability model for all of the parametric situations. It may also happen that the power is higher for some probability model other than the exponential, even when the exponential SSS has been used for the cluster detection.[18]

Table 4.1: Simulation study results for the exponential spatial scan statistic. Four probability models each with three different means inside true cluster are used under three right censoring cases: a=20%:20%, b=20%:40%, c=40%:20%. Outside cluster: Mean=2; Variance=4(Exponential), 0.188(Weibull), 2(log-Normal), and 1(Gamma).

Data	**IC**		**Power**			**PI**			**LC**		
Distribution	**M**	**V**	**a**	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**
Exponential	10	100.0	0.930	0.952	0.566	0.876	1.000	0.312	0.124	0.000	0.688
	15	225.0	0.941	1.000	0.831	0.920	1.000	0.659	0.080	0.000	0.341
	20	400.0	1.000	0.989	0.941	1.000	1.000	0.710	0.000	0.000	0.290
Weibull	10	4.0	0.696	0.645	0.492	0.836	0.876	0.654	0.164	0.124	0.090
	15	10	0.846	0.753	0.623	0.344	1.000	0.344	0.000	0.000	0.656
	20	7.0	0.852	0.877	0.883	1.000	1.000	0.243	0.000	0.000	0.757
Log-Normal	10	4.0	1.000	0.921	0.403	0.554	0.545	0.214	0.464	0.455	0.666
	15	10.0	0.941	1.000	0.742	0.708	0.612	0.119	0.292	0.388	0.881
	20	17.0	0.994	0.962	0.887	0.720	0.694	0.490	0.280	0.306	0.510
Gamma	10	5.0	0.705	0.699	0.545	0.781	0.739	0.709	0.291	0.261	0.008
	15	7.5	0.800	0.793	0.640	0.936	1.000	0.954	0.064	0.000	0.046
	20	10.0	1.000	0.795	0.652	1.000	0.850	0.957	0.000	0.150	0.043

IC=Inside Cluster M=Mean V=Varaince PI=Perfect Identification

LC=Large Cluster Identification

4.3.2 Simulation Study for Weibull Spatial Scan Statistic

For the Weibull SSS (Table 4.2, Figures B.3 and B.4), the overall results for the power and all the proportions' performances of the datasets were less variable than the results of the exponential and log-Weibull SSS's.

The power values of detecting a potential cluster were between 0.589 and 1 for the 20%:20% censoring setting, ranged from 0.621 to 1 for the 20%:40% censoring ratio, and between 0.579 and 1 for the 40%:20% case. The proportions of perfectly detecting a true cluster were high for all three censoring situations across all of the datasets, being least for the exponential model with the 40%:20% censoring ratio. A few observed non-zero proportions of datasets generated under the exponential and gamma distributions, who did not identify the true cluster were between 0.016 and 0.431.

The power values increased as the difference between the means of inside and outside the cluster increased and similar effects were seen for the strength of detection of the true cluster. This study showed that the Weibull SSS had more similar results for the spatial cluster detection regardless of the probability model used for the data generation.

Table 4.2: Simulation study results for the Weibull spatial scan statistic. Four probability models each with three different means inside true cluster are used under three right censoring cases: a=20%:20%, b=20%:40%, c=40%:20%. Outside cluster: Mean=2; Variance=4(Exponential), 0.188(Weibull), 2(log-Normal), and 1(Gamma).

Data	**IC**		**Power**			**PI**			**LC**		
Distribution	**M**	**V**	**a**	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**
Exponential	10	100.0	0.589	0.621	0.579	0.631	0.454	0.357	0.300	0.530	0.601
	15	225.0	0.909	0.844	0.791	0.704	0.492	0.522	0.291	0.077	0.478
	20	400.0	1.000	1.000	0.983	0.982	0.949	0.593	0.018	0.000	0.391
Weibull	10	4.0	1.000	0.979	0.947	0.989	0.963	0.944	0.011	0.037	0.056
	15	10.0	1.000	1.000	1.000	1.000	1.000	0.981	0.000	0.000	0.019
	20	7.0	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000
Log-Normal	10	4.0	0.865	0.932	0.842	0.747	0.873	0.798	0.253	0.127	0.202
	15	10.0	0.951	0.964	0.823	0.962	0.919	0.850	0.038	0.081	0.150
	20	10.0	1.000	1.000	0.968	1.000	1.000	1.000	0.000	0.000	0.000
Gamma	10	5.0	0.946	0.833	0.724	0.812	0.753	0.843	0.188	0.067	0.157
	15	7.5	1.000	1.000	0.960	1.000	0.714	0.000	0.000	0.014	0.000
	20	10.0	1.000	1.000	1.000	1.000	0.654	1.000	0.000	0.346	0.000

IC=Inside Cluster M=Mean V=Varaince PI=Perfect Identification

LC=Large Cluster Identification

4.3.3 Simulation Study for Log-Weibull Spatial Scan Statistic

Using the log-Weibull SSS (Table 4.3, Figures B.5 and B.6), the results showed that the powers varied from 0.432 to 0.969 for the 20%:20% censoring, from 0.411 to 0.824 for the 20%:40% censoring situation, and ranged from 0.324 to 0.631 for the 40%:20% censoring case. Overall, the maximum power was seen when the data were generated under the Weibull distribution and the minimum power was observed for the datasets distributed with the gamma probability model.

The proportions of datasets perfectly identifying the true cluster were moderate for the log-Weibull SSS. They were between 0.238 and 0.670 for the 20%:20% case, ranged from 0.259 to 0.663 for the 20%:40% censoring ratio, and between 0.133 and 0.490 for the 40%:20% censoring setting, respectively. The datasets from the gamma distribution had the highest proportions of large cluster identification including the true cluster among all four probability models, being highest for the 40%:20% censoring case. There was a decrement found in the power and the strength of identification of the true cluster for each model, when comparing the 20%:20% to the 20%:40% and 40%:20% censoring cases.

Table 4.3: Simulation study results for the log-Weibull spatial scan statistic. Four probability models each with three different means inside true cluster are used under three right censoring cases: a=20%:20%, b=20%:40%, c=40%:20%. Outside cluster: Mean=2; Variance=4(Exponential), 0.188(Weibull), 2(log-Normal), and 1(Gamma).

Data	**IC**		**Power**			**PI**			**LC**		
Distribution	**M**	**V**	**a**	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**
Exponential	10	100.0	0.594	0.531	0.324	0.431	0.309	0.216	0.365	0.351	0.590
	15	225.0	0.601	0.634	0.595	0.541	0.360	0.206	0.416	0.386	0.567
	20	400.0	0.696	0.548	0.619	0.455	0.431	0.349	0.381	0.274	0.439
Weibull	10	4.0	0.748	0.643	0.548	0.519	0.451	0.385	0.285	0.312	0.318
	15	10.0	0.892	0.696	0.597	0.558	0.579	0.490	0.272	0.296	0.329
	20	7.0	0.969	0.824	0.631	0.649	0.561	0.445	0.077	0.323	0.269
Log-Normal	10	4.0	0.541	0.458	0.503	0.483	0.433	0.302	0.389	0.245	0.488
	15	10.0	0.792	0.502	0.615	0.538	0.663	0.322	0.461	0.139	0.227
	20	17.0	0.684	0.634	0.597	0.670	0.604	0.466	0.322	0.194	0.239
Gamma	10	5.0	0.432	0.411	0.400	0.238	0.259	0.133	0.418	0.481	0.514
	15	7.5	0.495	0.483	0.436	0.363	0.261	0.292	0.428	0.472	0.628
	20	10.0	0.503	0.469	0.397	0.470	0.307	0.333	0.354	0.394	0.547

IC=Inside Cluster M=Mean V=Varaince PI=Perfect Identification

LC=Large Cluster Identification

4.4 Summary

The results from the simulation studies showed that, the power of detecting the potential cluster was higher for the 20%:20% ratio as compared to the 20%:40% and 40%:20% settings. This comparison was also true in the context of identification of a true cluster. The effect of the right differential censoring on power values and proportions of detection of the true cluster was similar, irrespective of what distribution from the exponential, Weibull, and log-Weibull for the SSS was used.

The true cluster was constructed with 25 individuals inside the cluster and 475 individuals outside the cluster. The results from the right differential censoring situations make sense because as we lose more information from the smaller data (inside cluster), the power and strength of detection of a true cluster may diminish. More precisely, if 40% of the censoring is imposed on 25 individuals instead of 20%, then the chance of losing the power of capturing a true cluster may increase. Overall, the highest values of the power and proportions of datasets who detected the true cluster were observed for the balanced censoring, i.e., the 20%:20% censoring setting.

For all of the probability models under the three SSS's, as the difference between means of time to event data increased inside and outside the true cluster, the power and proportion of detection of the true cluster also increased. It can be observed from the overall results of the three SSS's that the Weibull SSS had good power for detecting a potential cluster for the datasets distributed with any of the four probability models used in this study. Also for the identification of the true cluster, the Weibull SSS showed less variability on the simulated datasets than the exponential and log-Weibull SSS's.

Chapter 5

Discussion and Future Work

The spatial scan statistic (SSS) is a widely used statistical technique for the identification of the spatial clusters of different data types by using various probability distributions. In the context of time to event data, the SSS has the ability to detect if there are potential geographical clusters of cases with either longer and/or shorter than expected event times. These clusters can be adjusted for censoring, if the appropriate probability model is used. The SSS's for the exponential[18] and Weibull[19] distributions have already been developed.

In this study, we have constructed the SSS for the log-Weibull distribution as an alternative approach for detecting spatial clusters for time to event data. The log-Weibull distribution, being a specialized case of the generalized extreme value distribution, has a wide application in extreme value theory for modeling extreme and rare events.

The new log-Weibull method and the exponential and Weibull SSS's were applied to administrative data from Alberta Health consisting of time from

ED discharge for an AFF presentation to 1^{st} specialist visit within 365 days in Alberta during 2010-2011. The specialist visit was defined as a visit to a physician specializing in cardiology or internal medicine for this study.

Results from the SSS's showed that the exponential, Weibull, and log-Weibull distributions have detected the same most likely cluster, i.e., the Peace Country, Northern Lights, and Aspen regional Health Authorities. The most likely cluster was comprised of the rural areas in northern Alberta which have sparse or low population. The results suggested that people living in these northern rural areas may not have regular or quick access to the follow-up care to a specialist after an ED presentation.

The simulation studies indicated that the SSS with a Weibull distribution has more power and strength of detecting the true cluster as compared to the exponential and log-Weibull distributions, when the random data were simulated from the exponential, Weibull, gamma, and log-Normal distributions. Under three different situations of right censoring imposed on the simulated datasets, the Weibull SSS's power and detection of the true cluster was the most similar across all of the datasets for different parametric situations inside and outside the zones.

There are many aspects that can be seen as future work arising from this study. First, the proposed methodology based on the SSS for the log-Weibull distribution does not adjust for important factors such as age and gender. In future, such covariates can be adjusted in the analysis of the identification of potential clusters for time to event data. Second, the new developed method can only be performed on a purely spatial setting. The space-time scan statistic has been developed by other authors in both retrospective[28] and prospective[29] ways. In the future, the SSS based on the log-Weibull distribution can be extended to the space-time setting, and similar simulation studies can be performed to investigate power and detection of true space-time clusters. Third, Prates et al. have evaluated the bias of estimated relative risks from the spatial and space-time scan statistics.[59] The same approach can be applied on the new developed method in this study for the investigation of bias of relative risk estimates of the detected clusters from both space and space-time scan statistics for time to event data.

In summary, we have provided a new SSS using the log-Weibull distribution, applied the new method to specialist follow-up data in Alberta, and compared and contrasted SSS's for time to event data on real and simulated data. The covariates' adjustment, extension to the space-time scan statistics, and the relative risk estimates of detected clusters can be considered as future work of this study.

Bibliography

1. California Department of Public Health, EHIB. http://www.ehib.org/topic.jsp. [Accessed August 4, 2015].

2. Naus JI. The distribution of the size of the maximum cluster of points on the line. *Journal of the American Statistical Association* 1965; **60**: 532-538.

3. Naus JI. Clustering of random points in two dimensions. *Biometrika* 1965; **52**: 263-267.

4. Jacquez GM, Kaufmann A, Meliker J, Goovaerts P, AvRuskin G, Nriagu J. Global, local and focused geographic clustering for case-control data with residential histories. *Environmental Health* 2005; **4**: 4.

5. What is kriging? http://www.kriging.com/whatiskriging.html. [Accessed September 09, 2015].

6. Montero JM, Fernãndez-Aviles G, Mateu J. *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging* 2015; Wiley: United Kingdom, ISBN: 978-1-118-41318-0.

7. Thomas AJ, Carlin BP. Late detection of breast and colorectal cancer in Minnesota counties: An application of spatial smoothing and clustering. *Statistics in Medicine* 2003; **22**: 113-127.

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

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

10. Petrisor AI, Drane JW, Dragomirescu L. Using kriging and the DAC statistic to predict low birthweight clusters in Spartanburg County, SC. *Joint Statistical Meetings-Section on Statistics & the Environment* 2002; 2687-2691.

11. Stevenson JR, Emrich CT, Mitchell JT, Cutter SL. Using building permits to monitor disaster recovery: A spatio-temporal case study of coastal Mississippi following Hurricane Katrina. *Cartography and Geographic Information Science* 2010; **37(1)**: 57-68.

12. Coulston JW, Riitters KH. Geographic analysis of forest health indicators using spatial scan statistics. *Environmental Management* 2003; **31**: 764-773.

13. Marcos RDLF, Marcos CDLF. From star complexes to the field: Open cluster families. *Astrophysical Journal* 2008; **672**: 342-351.

14. Usher BM, Allen KL. Identifying kinship clusters: SatScan for genetic spatial analysis. *American Journal of Physical Anthropology* 2005; **126(S40)**: 210-211.

15. Margai F, Henry N. A community-based assessment of learning disabilities using environmental and contextual risk factors. *Social Science and Medicine* 2003; **56**: 1073-1085.

16. Costa MA, Assunção RM. A fair comparison between the spatial scan and the Besag-Newell disease clustering tests. *Environmental and Ecological Statistics* 2005; **12**: 301-319.

17. Kulldorff M, Information Management Services I (2011). SaTScan^{TM} v9.1.1: Software for the spatial and space-time scan statistics.
http://www.satscan.org/.

18. Huang L, Kulldorff M, Gregorio D. A spatial scan statistic for survival data. *Biometrics* 2007; **63**: 109-118.

19. Bhatt V, Tiwari N. A spatial scan statistic for the survival data based on Weibull distribution. *Statistics in Medicine* 2013; **33**: 1867-1876.

20. Jung I, Kulldorff M, Klassen A. A spatial scan statistic for ordinal data. *Statistics in Medicine* 2007; **26**: 1594-1607.

21. Jung I, Kulldorff M, Richard OJ. A spatial scan statistic for multinomial data. *Statistics in Medicine* 2010; **29**: 1910-1918.

22. Kulldorff M, Huang L, Konty K. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 2009; **8**: 58.

23. Costa MA, Kulldorff M. Application of spatial scan statistics: A review. (2009). In Glaz J, Pozdnyakov V, Wallenstein S. (Eds.), *Scan Statistics: Methods and Applications, Statistics for Industry and Technology* (pp. 129-152). Birkhauser, Boston: Springer.

24. Kulldorff M. A spatial scan statistic. *Communication in Statistics-Theory and Methods* 1997; **26**: 1481-1496.

25. Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R. Multivariate spatial scan statistics for disease surveillance. *Statistics in Medicine* 2007; **26**: 1824-1833.

26. Huang L, Huang L, Tiwari R, Zuo J, Kulldorff M, Feuer E. Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association* 2009; **104**: 886-898.

27. Rosychuk RJ, Chang H-M. A spatial scan statistic for compound Poisson data. *Statistics in Medicine* 2013; **32**: 5106-5118.

28. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health* 1998; **88**: 1377-1380.

29. Kulldorff M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society* 2001; **A164**: 61-72.

30. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; **25**: 3929-3943.

31. Iyengar VS. On detecting space-time clusters. In Proceedings *10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2004: 587-592.

32. Takahashi K, Kulldorff M, Tango T, Yih K. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics* 2008; **7**: 14.

33. Torabi M, and Rosychuk RJ. An examination of five spatial disease clustering methodologies for the identification of childhood cancer clusters in Alberta, Canada. *Journal of Spatial and Spatio-Temporal Epidemiology* 2011; **2**: 321-330.

34. Neill DB. Fast subset scan for spatial pattern detection. *Journal of Royal Statistical Society* 2012; **74(2)**: 337-360.

35. Glaz J, Naus J, Wallenstein S. *Scan Statistics*. Springer-Verlag New York, Inc; 2001.

36. Reliablity HotWire. http://www.weibull.com/hotwire/issue56/relbasics56.htm. [Accessed September 16, 2015].

37. White JS. The moments of log-Weibull order statistics. *Technometrics* 1969; **11**: 373-386.

38. Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I. Fewer permutations, more accurate p-values. *Bioinformatics* 2009; **25**: i161-i168.

39. Abrams A, Kleinman K, Kulldorff M. Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics* 2010; **9(61)**: 61-72.

40. Alberta. https://en.wikipedia.org/wiki/Alberta. [Accessed July 7, 2015].

41. Statistics Canada. https://web.archive.org/web/20070210220013/http://www40.statcan.ca/l01/cst01/phys01.htm. [Accessed July 3, 2015].

42. Alberta Health Services. http://www.albertahealthservices.ca/5767.asp. [Accessed July 20, 2015].

43. Ellehoj E, Schopflocher D. Calculating small areas analysis: Definition of sub-regional geographic units in Alberta. Edmonton, Alberta: Alberta Health and Wellness; 2003.

44. Alberta Health Services: Annual report 2008-2009. http://www.albertahealthservices.ca/publications/ahs-pub-annual-report-2008-2009.pdf. [Accessed September 09, 2015].

45. Alberta Health Services: Annual report 2013-2014. http://www.albertahealthservices.ca/Publications/ahs-pub-2013-2014-annual-report.pdf. [Accessed September 09, 2015].

46. Rosychuk RJ, Mariathas HH, Garaham MM, Holroyd BR, Rowe, BH. Geographic clustering of emergency department presentations for atrial fibrillation and flutter in Alberta, Canada. *Academic Emergency Medicine* 2015; **22(8)**: 965-975.

47. Heart and Stroke Foundation. Heart disease - Atrial Fibrillation (2009). http://www.heartandstroke.com/site/c.ikIQLcMWJtE/b.2796497/k.F922/Heart_Disease_Stroke_and_Healthy_Living.htm. [Accessed August 27, 2015].

48. Atrial Fibrillation Health Center. http://www.webmd.com/heart-disease/atrial-fibrillation/atrial-flutter. [Accessed July 20, 2015].

49. Friberg J, Buch P, Scharling H, Gadsbphioll N, Jensen GB. Rising rates of hospital admissions for atrial fibrillation. *Epidemiology* 2003; **14(6)**: 666-672.

50. Wattigney WA, Mensah GA, Croft JB. Increasing trends in hospitalization for atrial fibrillation in the United States, 1985 through 1999: Implications for primary prevention. *Circulation* 2003; **108(6)**: 711-716.

51. Alberta Health and Wellness. Ambulatory Care in Alberta using Ambulatory Care Classification System Data. Edmonton, Alberta: Alberta Health and Wellness; 2004.

52. Alberta Ambulatory Care Reporting Manual. http://www.health.alberta.ca/documents/ACRM-09-pt0-7.pdf. [Accessed July 22, 2015].

53. Canadian Institute for Health Information. https://www.cihi.ca/en/nacrs_exec_summ_2010_2011_en.pdf. [Accessed September 10, 2015].

54. Canadian Institute for Health Information. https://www.cihi.ca/en/nacrs_faq_jan_2012_en.pdf. [Accessed September 10, 2015].

55. Government of Alberta: Comparison of Alberta population counts between the AHCIP and the 2006 census. http://www.health.alberta.ca/documents/Population-2006-Comparison-2009.pdf. [Accessed August 07, 2015].

56. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

57. TIBCO Software Inc. S-PLUS 8 Version 8.1.1. 2008.

58. Feinberg WM, Blackshear JL, Laupacis A, Kronmal R, Hart RG. Prevalence, age distribution, and gender of patients with atrial fibrillation: analysis and implication. *Archives of International Medicine* 1995; **155(5)**: 469-473.

59. Prates MO, Kulldorff M, Assunção RM. Relative risk estimates from spatial and space-time scan statistics: Are they biased? *Statistics in Medicine* 2014; **33(15)**: 2634-2644.

Appendix A

Geographic Units

Table A.1: Regional Health Authority (RHA) and sub-Regional Health Authority (sRHA) codes and names.

RHA	Code	sRHA
R1 Chinook RHA	1	R101 Crowsnest Pincher Creek
	2	R102 Ft Mcleod Cardston
	3	R103 Lethbridge
	4	R104 Picture Butte Raymond Milk River
	5	R105 Vauxhall Taber
R2 Pallisor Health Region	6	R201 Palliser North and Central
	7	R202 Palliser West
R3 Calgary Health region	8	R301 Calgary North East
	9	R302 Calgary Beddington Heights
	10	R303 Calgary Northwest
	11	R304 Calgary University
	12	R305 Calgary Charleswood
	13	R306 Calgary Marlborough
	14	R307 Calgary Shaganappi
	15	R308 Calgary Bowness
	16	R309 Calgary Scarboro
	17	R310 Calgary Forest Lawn
	18	R311 Calgary Lakeview
	19	R312 Calgary Mount Royal
	20	R313 Calgary Haysboro
	21	R314 Calgary Bonavista
	22	R315 Calgary South
	23	R320 Banff-Canmore
	24	R321 Didsbury-Strathmore
	25	R322 Vulcan-Claresholm
	26	R323 High River-Black Diamond
R4 David Thompson RHA	27	R401 Clearwater
	28	R402 Brazeau
	29	R403 Wetaskiwin-Hobbema
	30	R404 Ponoka
	31	R405 Lacombe
	32	R406 Red Deer
	33	R407 Olds
	34	R408 Drumheller-Hanna
	35	R409 Stettler-Consort
R5 East Central Health	36	R501 Region 5 Northwest
	37	R502 Regions 5 Northeast
	38	R503 Region 5 Southeast
	39	R504 Region 5 South Central
	40	R505 Region 5 Southwest

RHA	Code	sRHA
R6 Capital Health	41	R601 St. Albert
	42	R602 Edmonton Castle Downs
	43	R603 Edmonton Woodcroft
	44	R604 Edmonton Eastwood
	45	R605 Edmonton North Central
	46	R606 Edmonton North East
	47	R607 Edmonton Bonnie Doon
	48	R608 Edmonton West Jasper Place
	49	R609 Edmonton Twin Brooks
	50	R612 Edmonton Mill Woods
	51	R613 Sherwood Park
	52	R614 Strathcona County
	53	R615 Thorsby
	54	R616 Leduc Office
	55	R617 Beaumont
	56	R618 Westview
	57	R619 Sturgeon County
	58	R620 Fort Saskatchewan
R7 Aspen RHA	59	R701 Aspen West
	60	R702 Aspen Central
	61	R703 Aspen North
	62	R704 Aspen East
R8 Peace Country Health	63	R801 Peace NW
	64	R802 Peace NE
	65	R803 Peace SE
	66	R804 Peace SW
R9 Northern Lights Health Region	67	R901 High Level
	68	R902 La Crete
	69	R903 Northern Lights NW
	70	R904 Fort McMurray

Table A.2: Nearest neighbours for each sub-Regional Health Authority leading to the combined population of $\leq 10\%$ of the total population for the feasibility of the analysis and time restrictions.

sRHA	**Nearest Neighbours**
R101	1 2 26 3 22
R102	2 3 4 1 5 25 26 21
R103	3 4 2 5 25 1 26 7
R104	4 3 5 2 25 7 1 6 26
R105	5 4 3 7 6 2 25 1 26
R201	6 7 5 4 3 25 34 35 2
R202	7 5 6 25 34 4 3 35
R301	8 11 15 9
R302	9 12 8 10
R303	10 13 12 9 17
R304	11 8 15 14
R305	12 9 15 10
R306	13 17 10 16 12
R307	14 15 11 18
R308	15 14 11 12 18
R309	16 17 19 20 13
R310	17 16 13 20 19
R311	18 19 20 15 14
R312	19 18 20 16 21
R313	20 19 21 16 18
R314	21 20 22 19
R315	22 21 20 19
R320	23 27 8 33 11 14
R321	24 10 13 9
R322	25 26 21 17 13
R323	26 22 21 20
R401	27 28 23 33 29 56 30 59 53 31
R402	28 56 53 29 30 48 60
R403	29 30 53 31 54 49 48
R404	30 29 31 53 32 54 55
R405	31 32 30 29 33 53 54 40 55
R406	32 31 33 30 29 24
R407	33 32 8 9
R408	34 35 7 24 25 32
R409	35 39 38 34 40 36 37 7 32 31
R501	36 39 62 52 40 58 55 51 46
R502	37 38 36 62 39 35 40 52 58 55 51
R503	38 37 35 39 36 40 62 34 52 55 58 51
R504	39 40 36 35 38 55 52 37 54 51
R505	40 55 39 54 50 52 51

sRHA	Nearest Neighbours
R601	41 42 43 45
R602	42 41 45 43 44
R603	43 41 42 48
R604	44 45 47 42
R605	45 44 42 46
R606	46 51 45 44
R607	47 44 50 45
R608	48 43 49
R609	49 48 47
R612	50 47 49
R613	51 46 47 44
R614	52 58 51 46 45 44
R615	53 29 48 49 54
R616	54 55 50 49
R617	55 54 50 51 47
R618	56 28 53 60 48 29
R619	57 46 45 42 58 41
R620	58 52 46 51 57 45
R701	59 65 27 66 28 56 60 53 29
R702	60 56 41 57 28 43
R703	61 70 62 57 58 69 60 52 46
R704	62 36 37 52 58 39 46 57 51
R801	63 67 64 66 65 68 59 60 56 28
R802	64 65 68 63 60 67 66 61 59 56
R803	65 64 59 66 60 63 56 28 53 27
R804	66 65 59 63 64 60 27 28 56 67 53
R901	67 63 68 64 65 66 69 59 61 60 56
R902	68 64 67 63 69 65 61 60 70 66
R903	69 70 61 68 62 64 60 57 58 36 52
R904	70 61 69 62 37 36 58 52 57

Table A.3: Total number of uncensored (complete) and censored observations per sub-Regional Health Authority.

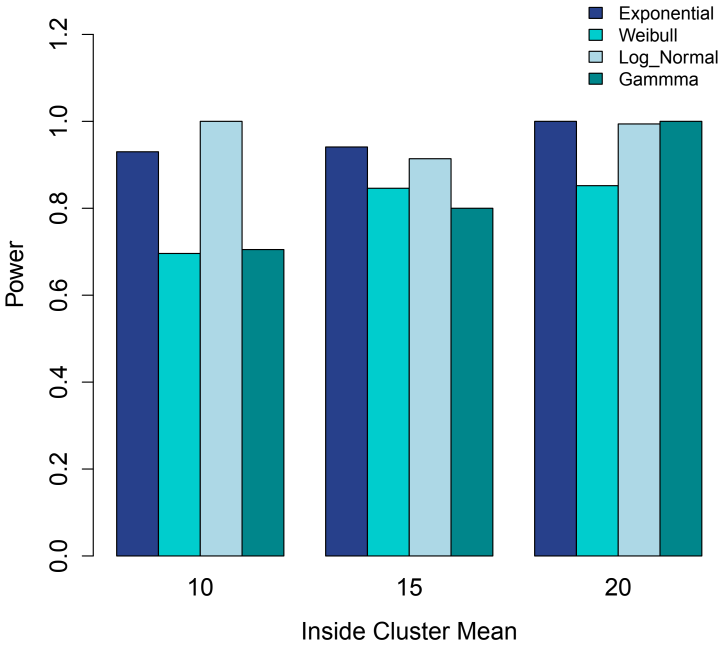
SRHA	Uncensored	Censored
R101	11	16
R102	29	15
R103	102	18
R104	37	11
R105	14	3
R201	87	26
R202	15	6
R301	50	6
R302	26	4
R303	34	0
R304	38	5
R305	64	7
R306	26	5
R307	63	7
R308	47	15
R309	18	2
R310	32	11
R311	44	8
R312	45	7
R313	66	10
R314	63	7
R315	43	13
R320	37	11
R321	94	25
R322	18	7
R323	58	23
R401	25	10
R402	10	1
R403	27	11
R404	24	7
R405	23	6
R406	93	18
R407	33	12
R408	33	35
R409	19	10
R501	25	11
R502	6	1
R503	14	8
R504	19	6
R505	23	10

SRHA	**Uncensored**	**Censored**
R601	59	10
R602	43	2
R603	79	9
R604	39	9
R605	65	7
R606	59	7
R607	92	12
R608	106	17
R609	94	5
R612	74	6
R613	54	2
R614	20	4
R615	3	0
R616	21	2
R617	11	2
R618	96	17
R619	21	2
R620	17	7
R701	37	13
R702	46	10
R703	32	17
R704	56	23
R801	26	29
R802	11	10
R803	29	16
R804	69	20
R901	6	2
R902	5	3
R903	11	5
R904	24	5

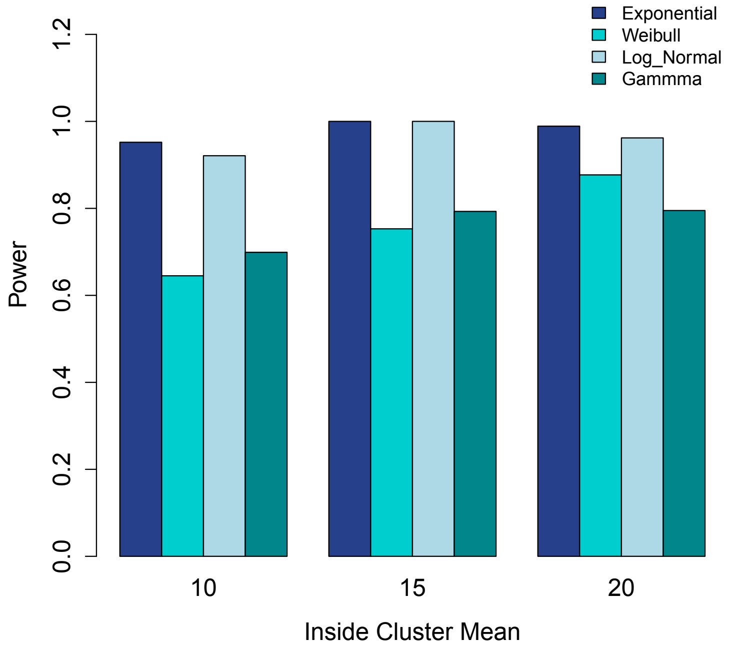
Appendix B

Figures for Simulation Study

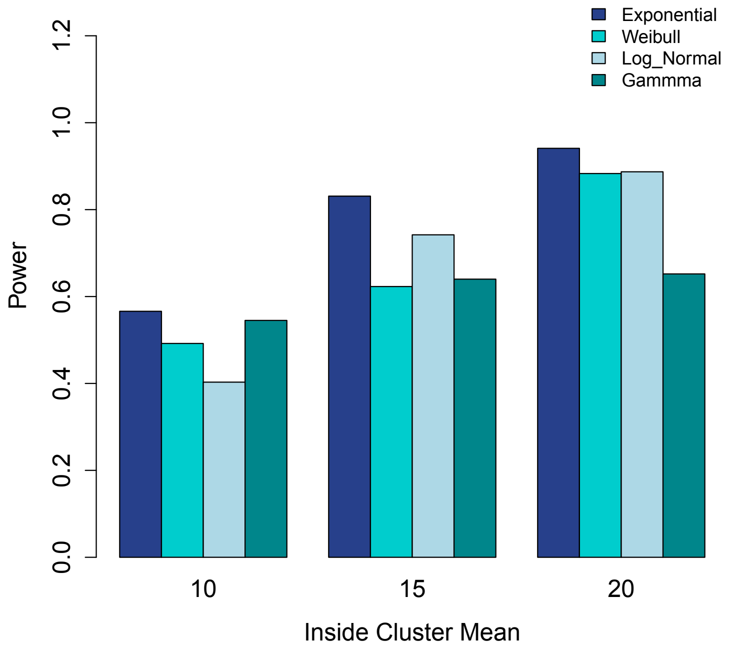
Figure B.1: Power of the exponential SSS for detecting a potential cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2.



(a) Censoring inside:outside cluster=20%:20%

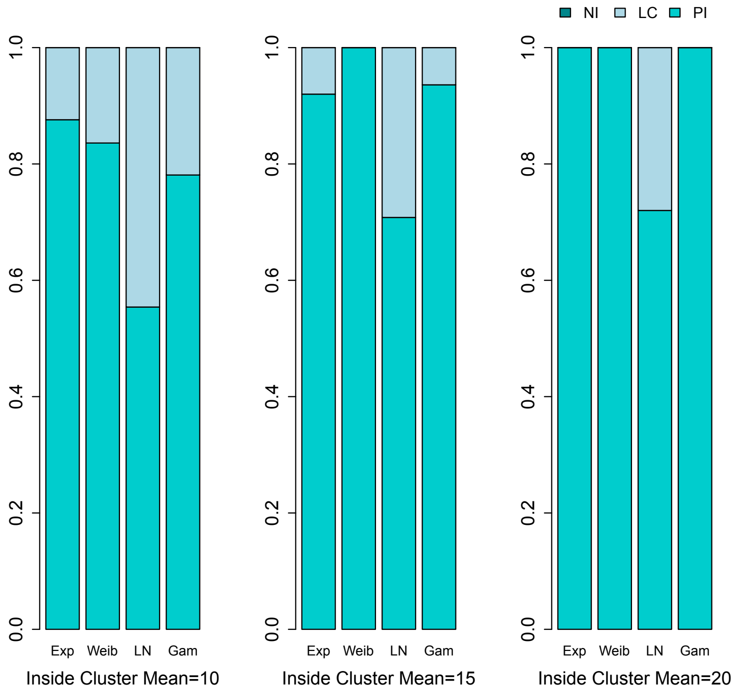


(b) Censoring inside:outside cluster=20%:40%

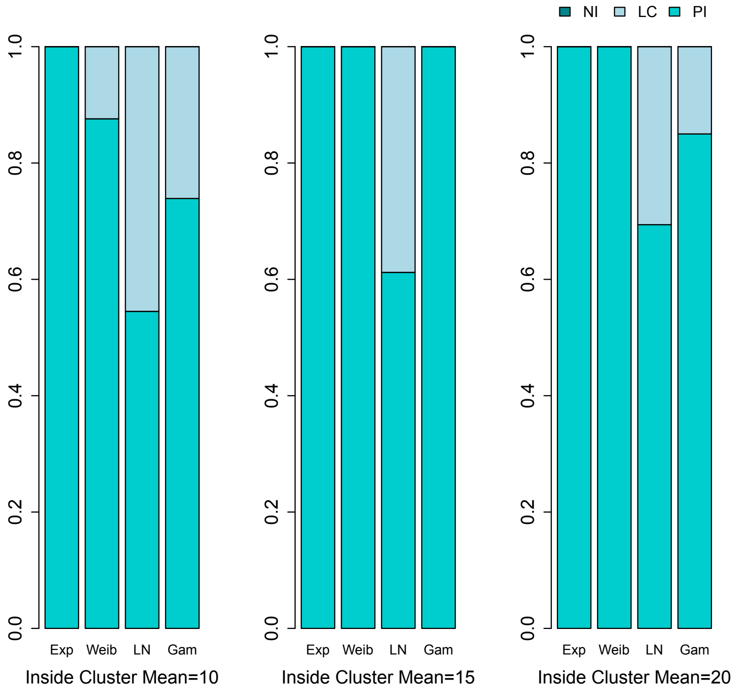


(c) Censoring inside:outside cluster=40%:20%

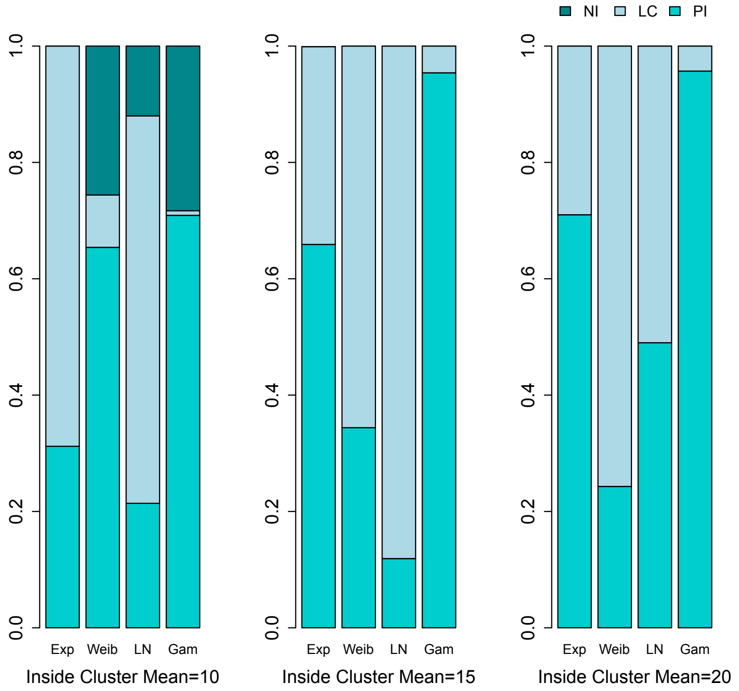
Figure B.2: Strength of the exponential SSS for detecting a true cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2. PI= Perfect Identification, LC= Large Cluster(including true), NI= No Identification.
Exp=Exponential, Weib=Weibull, LN=Log-Normal, Gam=Gamma.



(a) Censoring inside:outside cluster=20%:20%

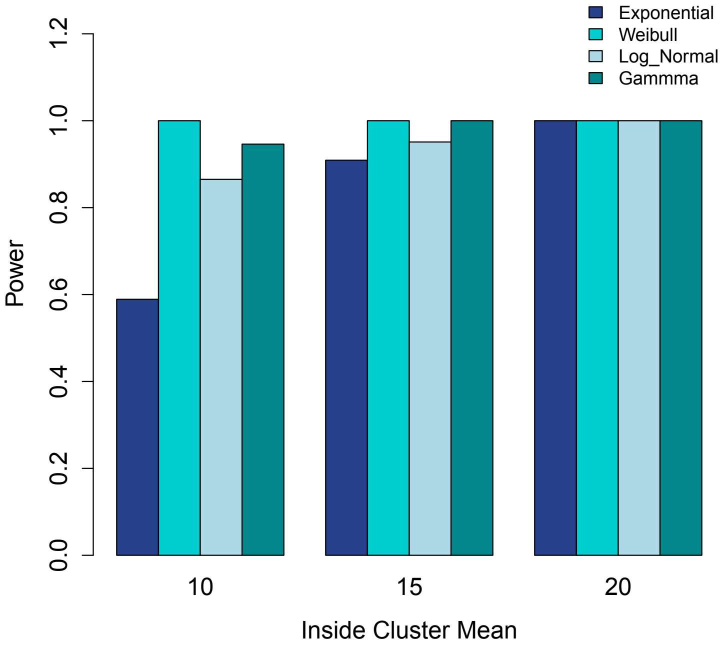


(b) Censoring inside:outside cluster=20%:40%

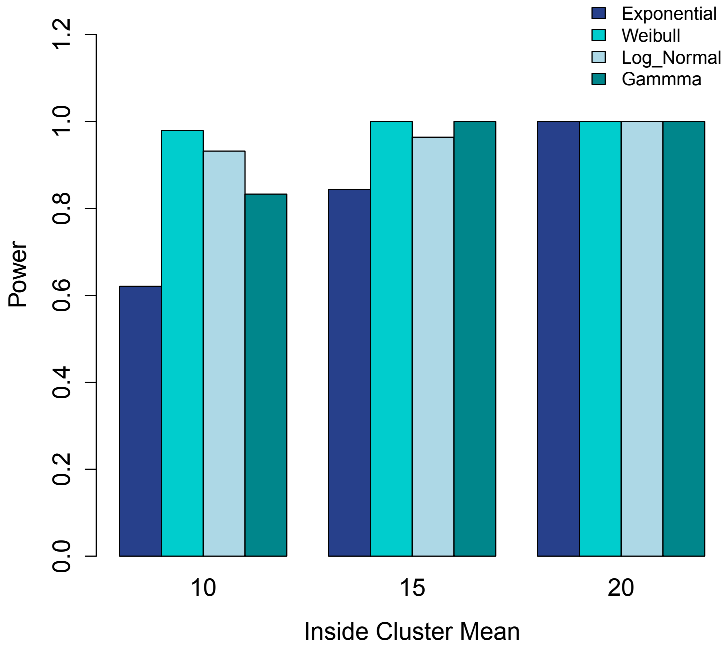


(c) Censoring inside:outside cluster=40%:20%

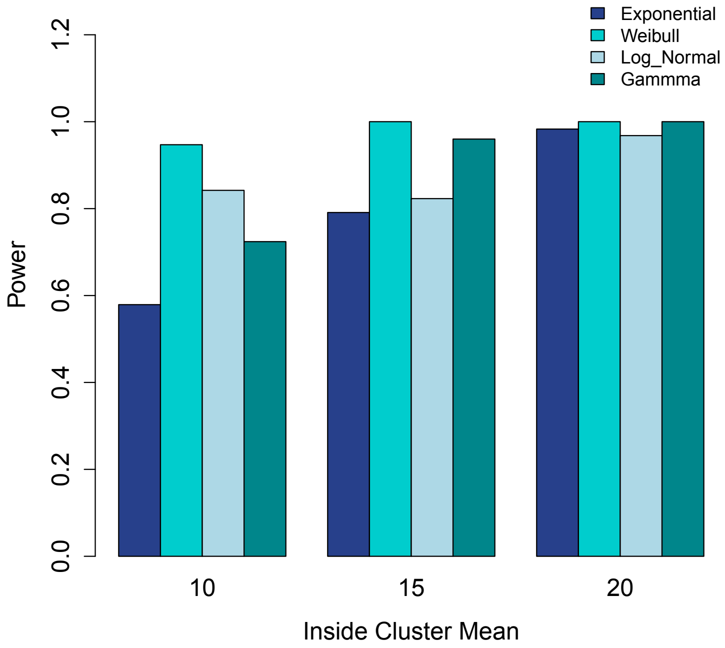
Figure B.3: Power of the Weibull SSS for detecting a potential cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2.



(a) Censoring inside:outside cluster=20%:20%

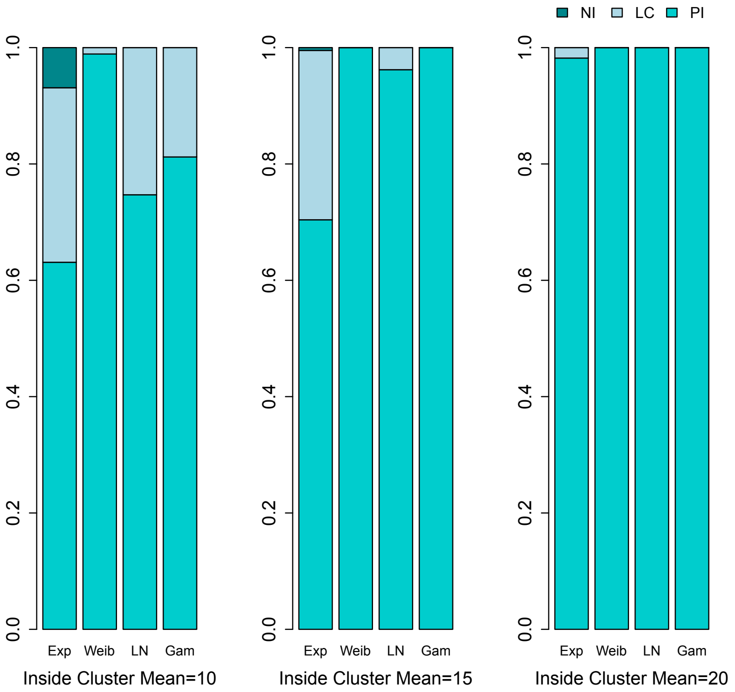


(b) Censoring inside:outside cluster=20%:40%

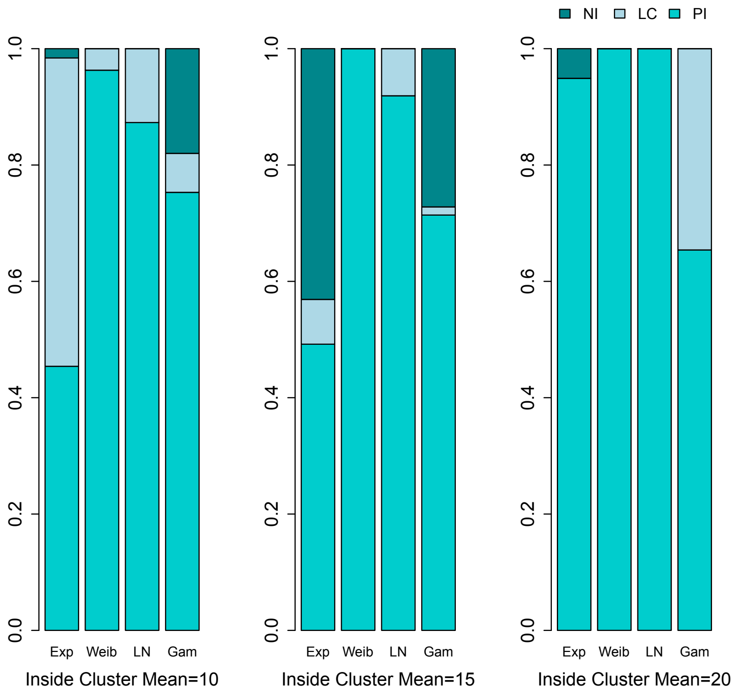


(c) Censoring inside:outside cluster=40%:20%

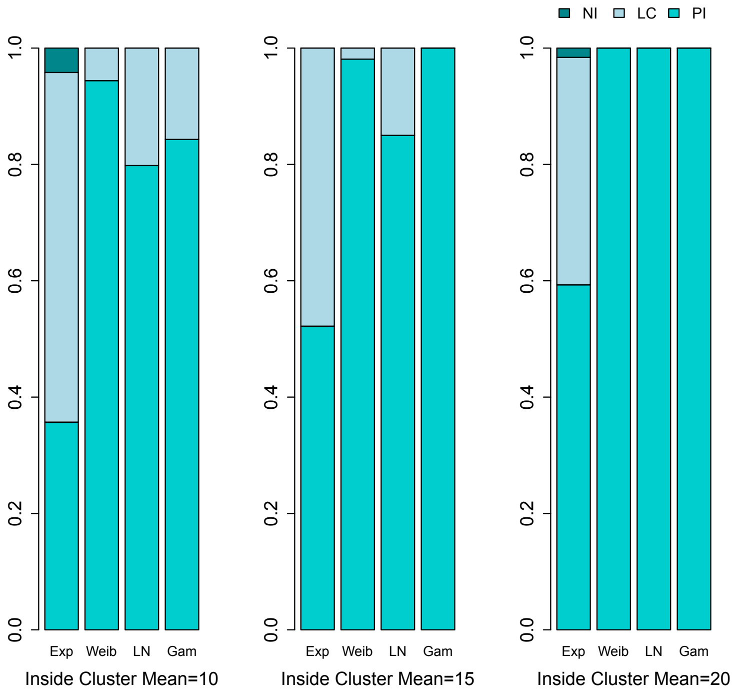
Figure B.4: Strength of the Weibull SSS for detecting a true cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2. PI= Perfect Identification, LC= Large Cluster(including true), NI= No Identification.
Exp=Exponential, Weib=Weibull, LN=Log-Normal, Gam=Gamma.



(a) Censoring inside:outside cluster=20%:20%

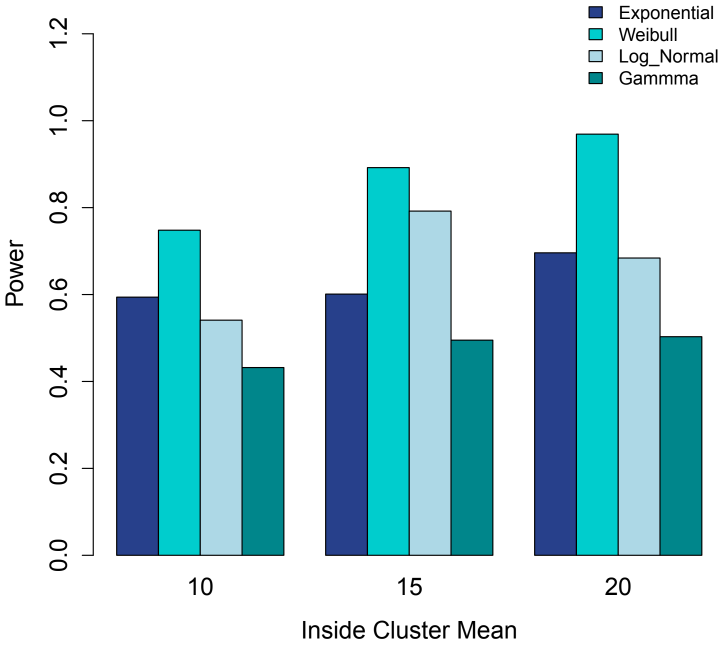


(b) Censoring inside:outside cluster=20%:40%

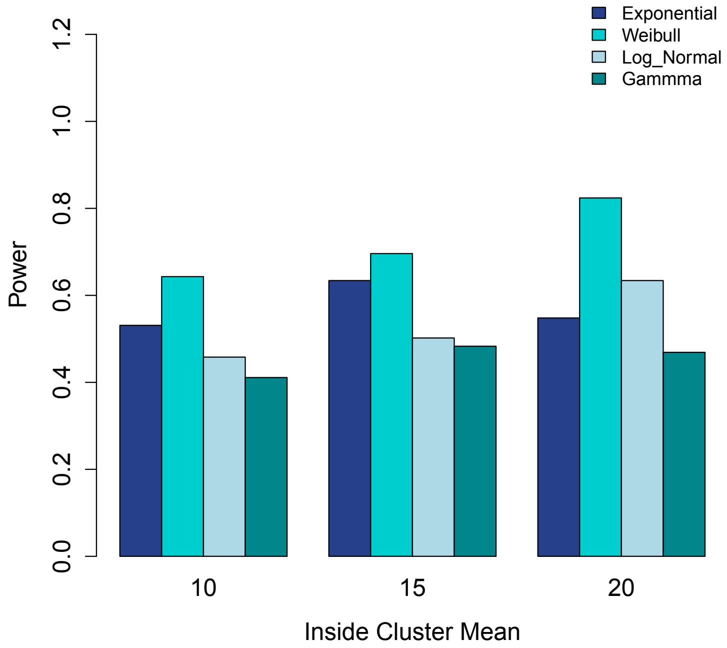


(c) Censoring inside:outside cluster=40%:20%

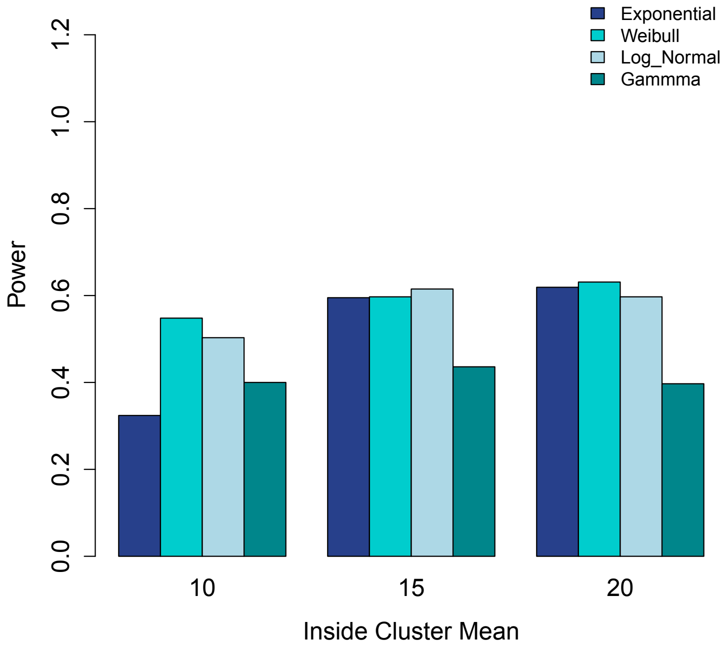
Figure B.5: Power of the log-Weibull SSS for detecting a potential cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2.



(a) Censoring inside:outside cluster=20%:20%

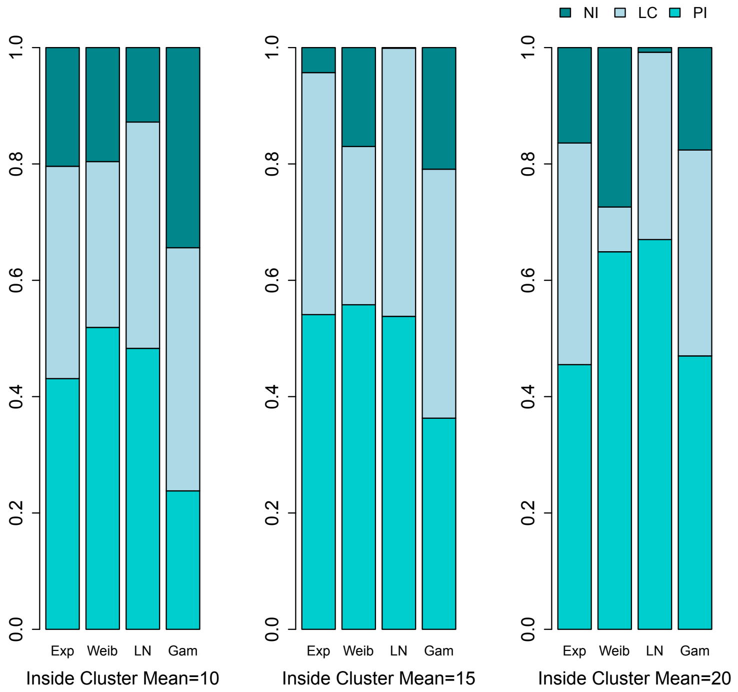


(b) Censoring inside:outside cluster=20%:40%

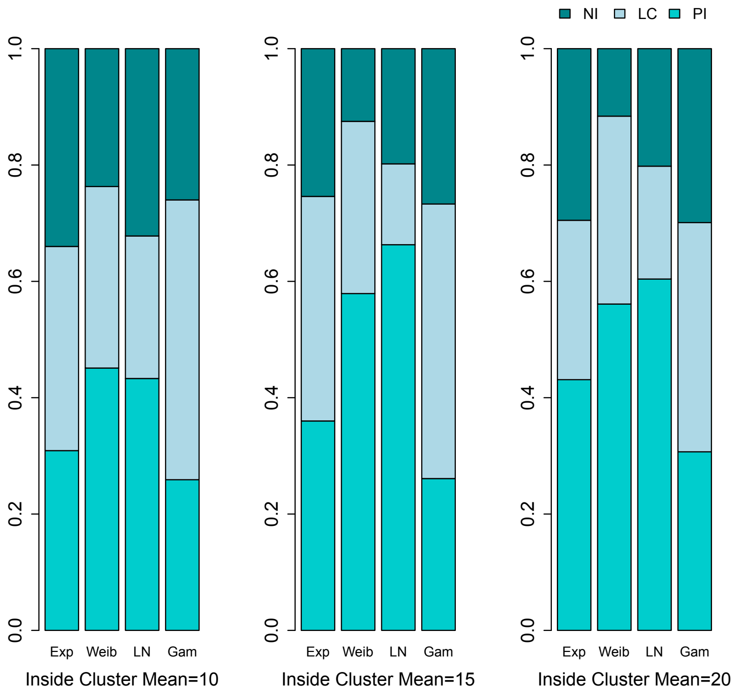


(c) Censoring inside:outside cluster=40%:20%

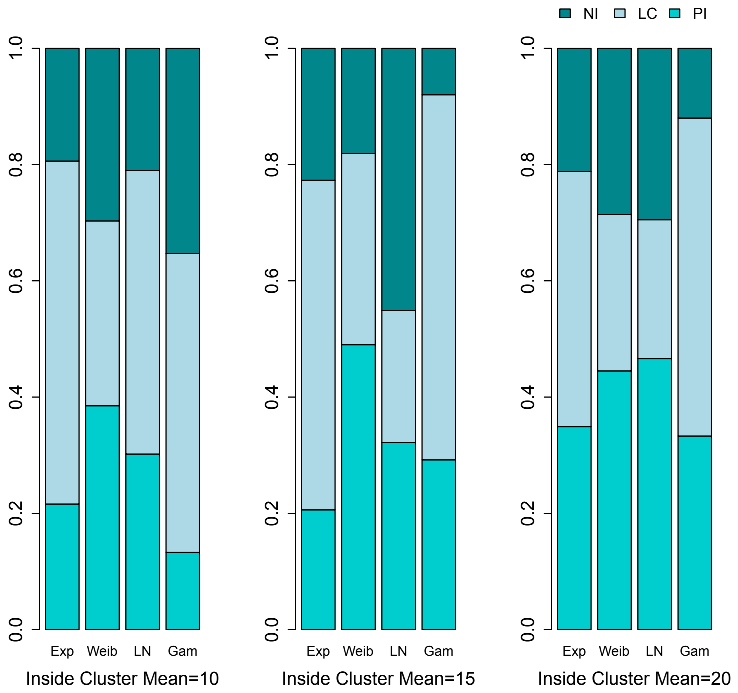
Figure B.6: Strength of the log-Weibull SSS for detecting a true cluster under right differential censoring. Datasets are generated using four probability models with outside cluster mean=2. PI= Perfect Identification, LC= Large Cluster(including true), NI= No Identification.
Exp=Exponential, Weib=Weibull, LN=Log-Normal, Gam=Gamma.



(a) Censoring inside:outside cluster=20%:20%



(b) Censoring inside:outside cluster=20%:40%



(c) Censoring inside:outside cluster=40%:20%